Continuous-Time Modeling of Counterfactual Outcomes Using Neural Controlled Differential Equations

Nabeel Seedat [* 1]   Fergus Imrie [* 2]   Alexis Bellot [3]   Zhaozhi Qian [1]   Mihaela van der Schaar [1 2 4]

Abstract

Estimating counterfactual outcomes over time has the potential to unlock personalized healthcare by assisting decision-makers to answer "what-if" questions. Existing causal inference approaches typically consider regular, discrete-time intervals between observations and treatment decisions and hence are unable to naturally model irregularly sampled data, which is the common setting in practice. To handle arbitrary observation patterns, we interpret the data as samples from an underlying continuous-time process and propose to model its latent trajectory explicitly using the mathematics of controlled differential equations. This leads to a new approach, the Treatment Effect Neural Controlled Differential Equation (TE-CDE), that allows the potential outcomes to be evaluated at any time point. In addition, adversarial training is used to adjust for time-dependent confounding which is critical in longitudinal settings and is an added challenge not encountered in conventional time-series. To assess solutions to this problem, we propose a controllable simulation environment based on a model of tumor growth for a range of scenarios with irregular sampling reflective of a variety of clinical scenarios. TE-CDE consistently outperforms existing approaches in all simulated scenarios with irregular sampling.

1. Introduction

Decision-makers must answer several critical questions before taking an action. In the clinical setting, before a treatment is given, clinicians must evaluate whether a treatment should be given and, if so, both *what* treatment is best for their patient and *when* the treatment should be administered. Answering such questions requires reliably estimating the effect of a treatment or sequence of treatments. While from a causal inference perspective, clinical trials represent the gold standard to answer these questions, it is highly desirable to estimate treatment effects from observational data. This is due to the significant expense, relatively small sample sizes, and narrow inclusion criteria of clinical trials.

There are several causal inference methods proposed in the static setting (e.g. Shalit et al., 2017; Alaa & van der Schaar, 2017; Yoon et al., 2018). However, estimating the effects of treatments *over time* is of paramount importance for real-world administration of complex treatment plans and personalized healthcare. Only in the longitudinal setting can we understand how diseases evolve under different treatment plans, how individual patients respond to treatment over time, or the optimal timing for treatment.

However, estimating counterfactual outcomes in the longitudinal setting introduces additional challenges, the most significant of which is that the observed treatment assignment may depend on confounding variables that vary over time (time-dependent confounding, Platt et al., 2009). For example, not all cancer patients are equally likely to be offered the same chemotherapy regimen. In particular, the history of patients' covariates and their response to past treatments affects future treatments (Bica et al., 2021). This can introduce bias in causal effects and variance in the estimation of counterfactuals due to the systematic differences in the distribution of confounding variables between any two sets of treatments over time.

This issue of time-dependent confounding and distribution shift is the primary challenge of causal inference over time, not encountered in standard time-series. Hence, conventional time-series models are not applicable to our setting as they do not adjust for bias introduced by time-varying confounders and hence are sensitive to the policy in the observational data (Schulam & Saria, 2017).

While prior work in causal inference has sought to mitigate such confounding bias (Robins et al., 2000; Lim et al., 2018; Bica et al., 2020b), the setting considered is overly restric-

---
[*]Equal contribution [1]Department of Applied Mathematics and Theoretical Physics, University of Cambridge, UK [2]University of California, Los Angeles, USA [3]Columbia University, USA [4]The Alan Turing Institute, London, UK. Correspondence to: Nabeel Seedat <ns741@cam.ac.uk>.
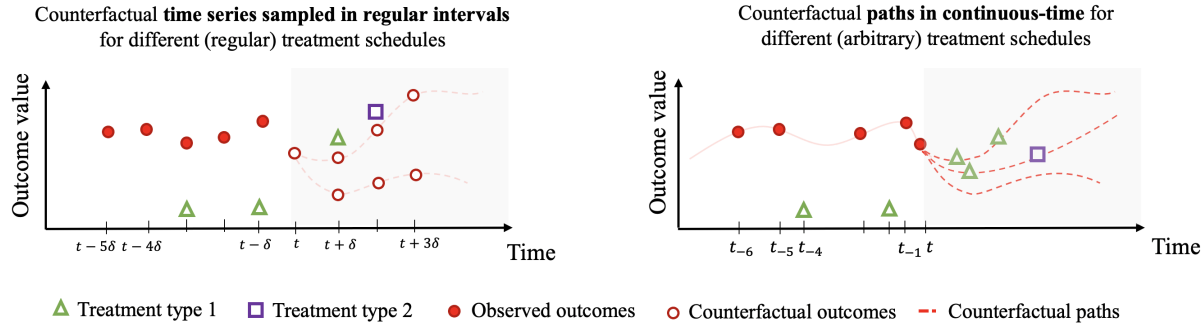
Figure 1: Illustration of the different paradigms of longitudinal data processing. We contrast the regular sampled setting (*left*) which RNN-based methods assume vs the irregularly sampled setting (*right*) which TE-CDE addresses, where data can be observed and evaluations carried out at any time-step.

tive and does not reflect most real-world observation data. In particular, previous work assumes that data is regular and arrives at fixed, evenly spaced time intervals and that the sampling times perfectly coincide between different individuals. However, neither is true in practice, significantly limiting the practical use of such methods.

Discretizing the patient's evolution over time, an inherently continuous process, has significant limitations, both when learning from historical data and for prospective clinical use. From a learning perspective, observational data is typically not sampled regularly. Indeed, irregularity in observational data can manifest for simple reasons, such as scheduling, a patient missing an appointment, or a healthcare practitioner not capturing the observation, to more complex considerations, for example more severe cases are often observed more frequently while different treatments can require differences in monitoring.

Prospective use cases raise similar issues surrounding mismatches between the discretization scheme and desired evaluation times that means the chosen discretization may not be applicable. As a result, for real-world applications where data is sampled irregularly, we believe that treatment effects over time should be modeled in a continuous manner.

Contributions. In this paper, we address the realistic but understudied problem of counterfactual estimation in the irregularly sampled setting with time-dependent confounding; a significantly more complex setting for counterfactual estimation than the standard regular, discrete setting.

To do so, we depart from existing methods based on recurrent neural networks (RNNs) and propose a novel alternative inspired by recent breakthroughs in neural controlled differential equations (CDEs) (Kidger et al., 2020), which we call the *Treatment Effect Neural Controlled Differential Equation (TE-CDE)*.

To model the observation histories, we learn a *continuous*

latent representation of the patient state as the solution to a CDE. To the best of our knowledge, this is the first work to frame the evolution of a patient's latent state as the solution to a CDE. This framing enables TE-CDE to learn from arbitrary historical observation patterns and allows potential outcomes to be evaluated at any point in time.

In addition, we introduce a controllable simulation environment based on a realistic model for tumor growth to generate irregularly sampled observational data. We demonstrate that the unrealistic assumptions imposed by existing state-of-the-art models lead to reduced performance in a range of irregularly sampled scenarios, and that TE-CDE outperforms these methods across all scenarios with irregularly sampled observation histories.

2. Related Work

This paper primarily engages with the literature on treatment effect estimation with time-varying covariates, treatments, and outcomes, but also draws on insights from causality in dynamical systems and recent work on modeling controlled differential equations. We explicitly note the difference between causal inference over time and conventional time series modeling as outlined in Section 1 and hence do not focus on recent advances in time series models. An extended discussion of related work can be found in Appendix B. In Table 1, we contrast the problem setting and assumptions of TE-CDE to other related work.

We argue for modeling the underlying continuous-time processes that give rise to the discrete observational data, which may itself be highly irregular. We contrast this approach with discrete-time methods that use a common discretization for all time series and are forced to interpolate and impute before model fitting. These methods also differ by how they adjust for confounding and for differences in covariate distributions in different treatment regimes. Marginal Structural Models (MSMs) are linear in treatment and covariate effect,

Table 1: Comparison of problem setting and assumptions. TE-CDE allows irregularly observed data and treatment plans defined in continuous time. Its assumptions are the continuous-time generalization of the standard assumptions in causal inference. The notations are defined in Section 3.

	Observation time	Treatment Plan	Overlap
Static setting	1	$\{0, 1\}$	$P(A\|X) > 0$
CRN/RMSN	$1, 2, \ldots, t \in \mathbb{N}^+$	$\{t+1, \ldots, t+k\} \to \{0, 1\}$	$P(A_{t+1}\|\mathcal{F}_{i,t}) > 0$
TE-CDE	$t_{i,0}, \ldots, t_{i,m_i} \in \mathbb{R}^+$	$[t, t'] \to \{0, 1\}$	$\lambda(t, \mathcal{F}_{i,t}) > 0$

and create a pseudo-population using inverse probability of treatment weighting, such that the probability of treatment does not depend on the time-varying confounders and thus effectively controlling for confounding bias (Robins et al., 2000). Lim et al. (2018) proposed a semi-parametric alternative to MSMs using recurrent neural networks to estimate propensity weights. The Counterfactual Recurrent Network (CRN, Bica et al., 2020b) uses a similar architecture but instead uses adversarial training to balance differences in covariate distributions in different treatment regimes. However, both assume data to be regularly sampled and fully observed at all time points, which is unrealistic in practice.

Gaussian process-based approaches such as Schulam & Saria (2017) are applicable to longitudinal data and take a continuous-time approach but in contrast, make strong assumptions about the model structure that is dependent on a particular application and prior knowledge of the form of the processes involved. Closer to the proposed approach, neural ordinary differential equations (ODE, Chen et al., 2018; Rubanova et al., 2019) and extensions (Kidger et al., 2020; Morrill et al., 2021) have been considered for modeling irregular time series data. However, neural ODE type methods are conventional time series models, which do not account for issues such as time-dependent confounding. In the context of intervention modeling, Gwak et al. (2020) proposed to use separate ODEs for interventions and outcome processes. However, they did so for systems with deterministic dynamics without integrating time-varying covariates and without addressing confounding. As a result, their approach is not applicable to treatment effect estimation in healthcare. Related is also Bellot & van der Schaar (2021) that proposed to model treatment effects in continuous time in the context of synthetic controls; however, contrasting our setting where there could be interventions over time, they only consider a single intervention at a particular time point and the approach is not applicable more generally to address multiple treatments.

3. Problem Formulation

We consider n i.i.d. individuals over a study period $[0, T]$. Each individual is represented by a d-dimensional path

$\mathbf{X} : [0, T] \to \mathbb{R}^d$, that defines the trajectory of patient covariates over time (and can include static covariates defined to be constant over time), a treatment process $A : [0, T] \to \{0, 1\}$ is a discrete path indicating treatment at each time $t \in [0, T]$, i.e. $A_t = a$, where $a \in \{0, 1\}$ and a counting process $N : [0, T] \to \mathbb{N}$ to denote the treatment assignment pattern of a single treatment over time, e.g. the number of treatments administered up to a given time[1]. These processes are assumed to control or modulate an outcome of interest $Y : [0, T] \to \mathbb{R}$, e.g. the tumor size of cancer patients over time, and we will distinguish between potential outcomes of Y, denoted $Y(A = a)$ or $Y(a)$ for simplicity, to define the potential outcome trajectory of patient i had it been given a treatment path defined by $A = a$.

In the context of electronic health records (EHRs) and most practical applications, the latent paths $\mathbf{X}$ are only partially-observed through m irregular observations, $\{(t_0, \mathbf{X}_{t_0}), (t_1, \mathbf{X}_{t_1}), \ldots, (t_m, \mathbf{X}_{t_m})\}$, with each $t_j \in \mathbb{R}$ the timestamp of the observation $\mathbf{X}_{t_j} \in \mathbb{R}^d$.

To avoid notation clutter, we use the time subscript to refer to function evaluation. The same observations apply to paths A and Y. The case where each i-th patient observation sequence has its own m_i irregular time stamps $t_{i,0}, \ldots, t_{i,m_i}$, thus differences in sampling intensity within a patient's trajectory and between different patients can be considered without modification of any part of the exposition. Indeed, analyzing time series data with such a complex pattern of observation is the central motivation of this work. Let $\mathcal{F}_t$ denote the filtration that is generated by all the observable events for a given individual up to time t, including observations of $\mathbf{X}_s, A_s$ and Y_s for $s \leq t$.

Our goal is to derive unbiased estimates of the potential outcomes at a given time t': $\mathbb{E}[Y_{t'}(A = a)|\mathcal{F}_t]$, for any value of time in the future $t' > t$, hypothesized discrete treatment path $A : [t, t'] \to \{0, 1\}$ with values a, given past observations up to time t, $\mathcal{F}_t$. However, with observational data only

---

[1]The definition can be generalized if multiple treatment types are considered, where N (and A) can be a multivariate process, each element of N counts a type of treatment assignment and would have a corresponding multivariate intensity process. For simplicity, our exposition considers a single treatment only.

one of these potential outcomes trajectories is observed for each unit depending on the treatment assignment. We refer to the unobserved potential outcomes as counterfactuals.

Potential outcomes processes are identifiable with respect to the filtration generated by the observed data under the following three assumptions. These three assumptions are the standard causal inference assumptions.

Assumption 1 (Consistency). *For an observed treatment process $A = a$, the potential outcome under this treatment trajectory is the same as the factual outcome $Y(a) = Y$.*

Assumption 2 (Overlap). *The intensity process $\lambda(t|\mathcal{F}_t)$ is not deterministic given any filtration $\mathcal{F}_t$ and time point $t \in [0, T]$, i.e.*

$$0 < \lambda(t|\mathcal{F}_t) = \lim_{\delta t \to 0} \frac{P(A_{t+\delta t} - A_t = 1|\mathcal{F}_t)}{\delta t} < 1. \quad (1)$$

Overlap means that there is some positive probability of treatment assignment at any point along a patient's trajectory over the time interval. It can be understood as a direct extension to the more familiar overlap assumption in the static context, $0 < P(\text{Treatment} = 1|\mathbf{x}) < 1$.

The last assumption extends unconfoundedness, or strong ignorability given a patient's trajectory, to ensure that it is sufficient to condition on the observed trajectory up to time t to block all backdoor paths, i.e. spurious correlation not part of the direct causal effect of interest, to the potential outcome at any time in the future. Similar to Assumption 2, unconfoundedness has previously been extended to the continuous-time domain for stochastic processes by Lok (2008); Saarela & Liu (2016); Ryalen et al. (2019).

Assumption 3 (Continuous-time sequential randomization). *The intensity process $\lambda(t|\mathcal{F}_t)$ with respect to the filtration $\mathcal{F}_t$ is equal to the intensity process generated by the filtration $\mathcal{F}_t \cup \{\sigma(Y_s) : s > t\}$ that includes the σ-algebras generated by future outcomes $\{\sigma(Y_s) : s > t\}$.*

It is worth mentioning that the intensity process plays the same role as propensity scores in discrete-time models (Robins, 1997), modeling the switching of the treatment process. Assumption 3 can thus be thought of as formalizing sequential randomization in the continuous-time model by stating that the intensity process does not depend on future potential outcomes, i.e. the current information is enough to estimate counterfactuals in the future without bias.

4. Treatment Effect Controlled Differential Equation

TE-CDE frames the latent trajectory of a patient's state, as a response to a controlled differential equation (CDE), driven by covariate, treatment, and outcome processes (Fig. 2), which to the best of our knowledge is the first to do so.

This formulation using a CDE permits to account for information available at $t > 0$ (rather than just initial value $t = 0$). In particular, neural controlled differential equations (Kidger et al., 2020; Morrill et al., 2021) allow incoming information to modulate the dynamics. This ability is natural in a clinical setting, as not only can we model the continuous-time latent state evolution of a patient trajectory, but also we account for incoming data (e.g. treatment changes) that modulate the dynamics of the system.

We now present key components needed to facilitate the modeling of *counterfactual* outcomes in *continuous* time. Additional properties of TE-CDE are discussed in Appendix D. The key components are as follows:
(1) TE-CDE's encoder learns a representation that is defined continuously in time (i.e. a continuous latent path), rather than only at discrete time steps.
(2) The latent path trajectory evolves as a response of a Neural Controlled Differential Equation (CDE).
(3) Decoding and prediction are in continuous-time.
(4) TE-CDE uses domain adversarial training to learn a representation that adjusts for time-dependent confounding and hence is suitable for causal estimation.

Encoding the latent path Z. TE-CDE's encoder ingests historical observations $\mathcal{F}_t$ up to time t and learns a latent path $\mathbf{Z} : [t_0, t] \to \mathbb{R}^l$ continuously over time that will be designed to be both predictive of the factual outcomes and agnostic of the observed assigned treatment. An explicit continuous-time representation allows us to process measurements with arbitrary observation patterns. We assume the initial state of the path $\mathbf{Z}_{t_0}$ to be parameterized by a neural network $g_\eta : \mathbb{R}^{d+1+1} \to \mathbb{R}^l$ embeds the initial outcome, covariate and treatment observations into a l-dimensional latent state which can be expressed as the solution to a CDE,

$$\mathbf{Z}_{t_0} = g_\eta(\mathbf{X}_{t_0}, A_{t_0}, Y_{t_0}), \quad (2)$$

$$\mathbf{Z}_t = \mathbf{Z}_{t_0} + \int_{t_0}^{t} f_\theta(\mathbf{Z}_s) \frac{d[\mathbf{X}_s, A_s, Y_s]}{ds} ds, \quad (3)$$

for $t \in (t_0, T]$ which denotes the present time, up to which observations of all processes are available. The dynamics of potential outcomes when controlled by the covariate and treatment process take the form of a CDE (Lyons et al., 2007). Hence, the solution $\mathbf{Z}$ is said to be the response of a Neural CDE (Kidger et al., 2020) driven or controlled by the covariate, treatment and outcome processes (concatenated into a vector $[\mathbf{X}_t, A_t, Y_t] \in \mathbb{R}^{d+1+1}$). In this sense, Neural CDEs are a family of continuous-time models that explicitly define the latent vector field $f_\theta : \mathbb{R}^l \to \mathbb{R}^{(d+1+1) \times l}$ by a neural network parameterized by θ, and the dynamics are modulated by the values of an auxiliary path over time.

We computationally obtain the latent path up to t from $\mathcal{F}_t$ by
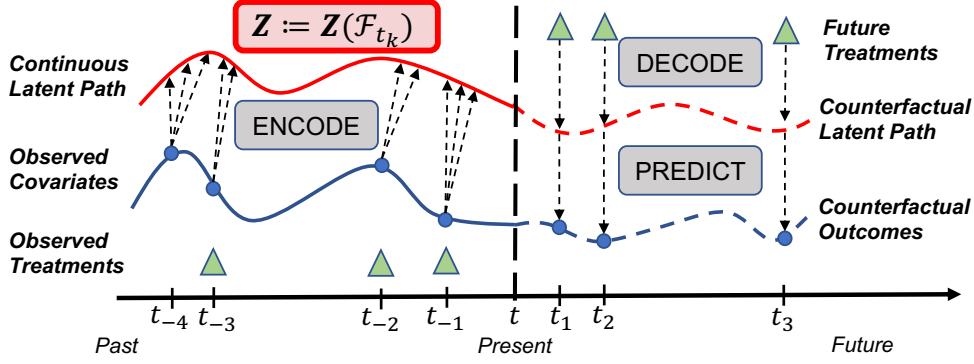
Figure 2: An illustration of TE-CDE. We learn a continuous latent path $\mathbf{Z}_t$ as the solution to a CDE by encoding historical observations. At future time points, we decode (hypothetical) future treatments to determine the latent state and use this to predict counterfactual outcomes.

solving the above initial value problem (IVP): $\forall s \in [t_0, t]$,

$$\mathbf{z}_s = \text{ODESolve}(f_\theta, \mathbf{Z}_{t_0}, \mathbf{X}_{\leq t}, A_{\leq t}, Y_{\leq t}), \quad (4)$$

where ODESolve denotes a numerical ODE solver as proposed by Kidger et al. (2020).

In practice, we have access to observations at certain (irregular) time points. Thus, we define an interpolation of the data with piece-wise continuous derivatives that serves as an approximation of the underlying paths[2].

Decoding and prediction. After the encoder processes all the observations up to the present time t, TE-CDE starts to decode and predict the potential outcomes up to some time $t' > t$ in the future for a hypothetical treatment schedule defined by the user. At this point, the latent path $\mathbf{Z}$ potentially changes as a result of the chosen treatment schedule, which can similarly be formalized using a second controlled differential equation such that,

$$\mathbf{Z}_{t'} = \mathbf{Z}_t + \int_t^{t'} f_\phi(\mathbf{Z}_s) \frac{dA_s}{ds} ds, \quad (5)$$

where t' denotes a desired time horizon, $\mathbf{Z}_t$ is the latent state of $\mathbf{Z}$ at time t which encodes the patient's history, and A_s represents the hypothetical treatment schedule for $t < s < t'$. $f_\phi : \mathbb{R}^{l+1} \to \mathbb{R}^l$ is a feed-forward neural network with trainable weights ϕ. As before, the decoded path can be obtained by solving the IVP:

$$\mathbf{Z}_s = \text{ODESolve}(f_\phi, \mathbf{Z}_t, A_{t \leq t'}). \quad (6)$$

Domain adversarial training for counterfactual estimation. The covariates $\mathbf{X}$ are time-dependent confounders,

which can increase variance in the estimation of counterfactuals if the treatment distribution is not properly balanced given a patient's trajectory (Mansournia et al., 2017). While unbiased by Assumption 3, counterfactual estimates may have lower variance given patient trajectories frequently observed in the data but higher variance for infrequently observed patient trajectories with consequences for performance generalization of the treatment effect as demonstrated by Shalit et al. (2017). To mitigate this confounding bias, we ensure the latent representation $\mathbf{Z}_t$ is *not* predictive of the observed treatment assignment pattern (Shalit et al., 2017; Bica et al., 2020b) which effectively induces representations that are balanced with respect to treatment assignment over time. The treatment invariance breaks the association between time-dependent confounders X_t and current treatment A_t.

At each time t, the j different treatments $A \in \{A_1, \ldots, A_j\}$ represent our domains. We then require at each timestep t, that the latent path Z_t be invariant across treatments options: $P(\mathbf{Z}_t | A_t = 0) = P(\mathbf{Z}_t | A_t = 1)$ and more generally equal across any two values in the domain of treatment options. In this context, distributions of the latent state differ across treatment groups if a classifier trained as a function $\mathbf{Z}_t$ to predict treatment assignment accurately separates the two groups. Such representations are called *balancing representation* as it balances the probability of the predicted treatment process $p(A_t = 1 | \mathbf{Z}_t) = 0.5$, i.e. minimizing the distributional variance between treatment groups in the representation space (Johansson et al., 2020).

We use two neural networks $h_\nu : \mathbb{R}^l \to \mathbb{R}^d$ and $h_a : \mathbb{R}^l \to [0, 1]$ to predict the outcome and treatment: $\forall s \in [t, t']$

$$\hat{y}_s = h_\nu(\mathbf{z}_s), \quad (7)$$

$$\hat{p}_s := \hat{p}(a_s = 1) = h_\alpha(\mathbf{z}_s). \quad (8)$$

Suppose there are $k \geq 1$ observations in the time window

---

[2]We note that since there may be discontinuities in $A_{\leq t}$ (e.g. treatment is applied in discrete stages) we can inform the solver about the jumps between pieces so that its integration steps may align with them. This can be achieved using the `jumpt` argument in the ODE solver (Chen et al., 2018; Kidger et al., 2020).

$[t, t']$ with observation times $(t_1, \ldots, t_k)$. The mean square error (MSE) of outcome prediction is defined as $\mathcal{L}^{(y)} = \frac{1}{k} \sum_{j=1}^{k} \left(y_{t_j} - \hat{y}_{t_j} \right)^2$. The cross entropy loss of treatment prediction is defined as: $\mathcal{L}^{(a)} = -\frac{1}{k} \sum_{j=1}^{k} a_{t_j} \log \left(\hat{p}_{t_j} \right) + (1 - a_{t_j}) \log \left(1 - \hat{p}_{t_j} \right)$.

We minimize the following loss function, enforcing simultaneous outcome prediction and balanced representations:

$$\mathcal{L} = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_i^{(y)} - \mu \mathcal{L}_i^{(a)}, \tag{9}$$

where $\mu > 0$ is a hyper-parameter controlling the trade-off between treatment and outcome prediction. Note that the minus sign before $\mathcal{L}^{(a)}$ would effectively *maximize* the treatment prediction loss and ensure that z_t is not predictive of treatment assignment A_t. This leads to balancing representations, which remove bias introduced by time-dependent confounders and allow for reliable counterfactual estimates.

Remarks on invariant representations. As shown by Johansson et al. (2019), invertible transformations (ϕ) are necessary for consistency of domain invariant representations (**Z**). We include for completeness that if ϕ is non-invertible there is information loss, which leads to unobservable error (η). Thus, we desire an invertible ϕ, which ensures $\eta = 0$. This highlights an important *strength of TE-CDE*, where by properties of ODEs/CDEs (Zhang et al., 2020), the representations from TE-CDE have *guaranteed invertibility*, since integration backward in time is always possible or we can alternatively integrate: $-f_\phi(\mathbf{Z}_s)$.

Intensity of sampling. It is well-known for EHR data that sampling frequency and observations (or lack thereof) carry information about the patient's health status (Alaa et al., 2017). In such cases, we can replace each observed tuple $(x_{t_j}, a_{t_j}, y_{t_j})$ with $(x_{t_j}, a_{t_j}, y_{t_j}, c_{t_j})$ where $c_{t_j} \in \mathbb{R}^{d+1+1}$ counts the number of times each one of the dimensions of X, A and Y have been observed up to time t_j. The extended tuple is fed into the encoder to inform it about the sampling.

5. Experiments

In this section, we validate the ability of TE-CDE to estimate counterfactual outcomes from irregularly sampled observational data. Since counterfactual outcomes are not known for real-world data, it is necessary to use synthetic or semi-synthetic data for empirical evaluation. First, we describe a simulation environment based on a Pharmacokinetic-Pharmacodynamic (PK-PD) model of lung cancer tumor growth (Geng et al., 2017), which allows counterfactuals to be calculated at any time point for arbitrary treatment plans. Furthermore, we introduce a continuous-time observation process based on Hawkes processes. The controllable nature of the observation process allows us to simulate irregularly

sampled observational data for a range of different observation process parameterizations, which are motivated by common healthcare scenarios.

5.1. Modeling tumor growth under general observation patterns

Tumor growth dynamics. We use a well-established biomathematical PK-PD model for tumor growth in lung cancer patients that includes the effects of chemotherapy and radiotherapy (Geng et al., 2017). The PK-PD model is representative of the true underlying physiological process with responses to interventions. Hence, results using the model should be closely representative of reality. Additionally, the same underlying model was also used by Lim et al. (2018) and Bica et al. (2020b). We briefly describe it below and refer the reader to Appendix C for more details. The tumor volume at time t after diagnosis is modeled as follows:

$$\frac{dV(t)}{dt} = \left(\underbrace{\rho \log \left(\frac{K}{V(t)} \right)}_{\text{Tumor growth}} - \underbrace{\beta_c C(t)}_{\text{Chemotherapy}} \right.$$
$$\left. - \underbrace{(\alpha_r d(t) + \beta_r d(t)^2)}_{\text{Radiotherapy}} + \underbrace{e_t}_{\text{Noise}} \right) V(t), \tag{10}$$

where chemotherapy concentration $C(t)$ and radiotherapy dose $d(t)$ are defined by their own equations (see Appendix C.1), $K, \beta_c, \alpha_r, \beta_r$ are effect parameters, and e_t accounts for randomness in tumor growth (Geng et al., 2017).

We consider four treatment options: no treatment, chemotherapy, radiotherapy, and combined chemotherapy and radiotherapy. The assignment of chemotherapy and radiotherapy are modeled as Bernoulli random variables with probabilities p_c and p_r, respectively, that depend on tumor diameter[3] as follows: $p_c(t) = \sigma \left(\frac{\gamma_c}{D_{\max}} \left(\bar{D}(t) - \theta_c \right) \right)$, $p_r(t) = \sigma \left(\frac{\gamma_r}{D_{\max}} \left(\bar{D}(t) - \theta_r \right) \right)$, where $D_{\max} = 13$cm is the maximum tumor diameter, $\theta_c = \theta_r = D_{\max}/2$ and $\bar{D}(t)$ is the average tumor diameter.

The degree of *time-dependent confounding* is controlled by γ_c and γ_r, where increasing $\gamma_{\{c,r\}}$ increases the probability that observational treatment assignment is based on tumor diameter (For more details see Appendix C.1).

Observation process. As discussed in Section 1, in real-world clinical settings, patients are rarely observed at fixed, regular time intervals. Instead, they are observed *irregularly*, with observations often a consequence of clinical factors, e.g. severity of illness, treatment regimen, or medical policy.

To simulate such nuances, we augment the simulation envi-

---

[3]Note we assume tumors are perfectly spherical to enable conversion between tumor volume and diameter.
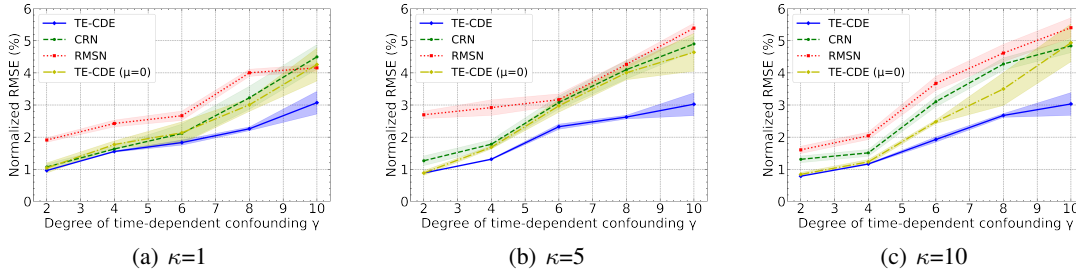
Figure 3: Counterfactual predictions for varying time-varying confounding γ. Additionally, various values of κ are illustrated which controls the intensity of sampling varying over a trajectory based on clinical stage.

ronment by modeling the patient observation process with a Hawkes process (Hawkes, 1971; Hawkes & Oakes, 1974). A Hawkes process is a flexible point process with temporal dependencies. Indeed, since clinicians are not memoryless and the times when they make observations often depend on past observations, the Hawkes process appears a sensible parameterization of such an observation process. This is especially true since it captures time-varying sampling intensity that depends on both patient history and clinical state.

These properties have been leveraged by both Bao et al. (2017) and Alaa et al. (2017), who used a Hawkes process to model different types of observational healthcare data, further justifying its applicability to healthcare and its use in our simulations. In addition, the success of both methods applying Hawkes processes to EHR data (which is often high-dimensional data), underlines the applicability to high-dimensional scenarios. From an experimental perspective, the Hawkes process is readily parameterized to simulate different clinical observation scenarios/regimes (i.e. a test bed to simulate different clinical scenarios).

Formally, the observation process $\{t_m\}_{m\in\mathbb{N}}$ is modeled as a univariate Hawkes process with a self-exciting intensity function and exponential kernel function (Lee et al., 2016):

$$\lambda(t, s_t = i) = \lambda_i^o + \sum_{\tau < t_m < t} e^{-2(t-t_m)},$$

$$\lambda_i^o = 0.01\kappa^i, \tag{11}$$

where s_t represents the clinical state of the patient at time t and τ represents the time of the previous change of state, i.e. $\tau = \max\{t' < t : s_t' = s_t,\ s_{t^-}' \neq s_t\}$. We define the clinical state using the American Joint Committee on Cancer staging system (Gershenwald et al., 2017). We convert the clinical stages $\{S1A, S1B, S2, S3\}$ to stages $s_t = \{0, 1, 2, 3\}$ (see Appendix C.2 for further details).

While in state $s_t = i$, the base observation intensity is given by $\lambda_i^o = 0.01\kappa^i$, where $\kappa \geq 1$ controls the change of base sampling intensity between cancer stages, resulting

in a higher base sampling intensity for patients with more advanced cancer, emulating real-world observation patterns.

Note that further control over the sampling process is possible, either by introducing additional states s_t or varying κ based on, for example, treatment. Examples of observation trajectories can be found in Appendix C.2.

Benchmarks. We compare TE-CDE [4,5] with state-of-the-art methods for counterfactual estimation over time CRN (Bica et al., 2020b) and RMSN (Lim et al., 2018) in different irregular sampling scenarios. Both CRN and RMSN rely on the assumption of regularly sampled data. Thus, for the irregular setting of interest, we divide the timeline equally, interpolate and impute the "un-observed" observations. We also evaluate a Gaussian Process (GP) based model for continuous-time, similar to Schulam & Saria (2017). However, the model performance is poor for larger values of time-dependent confounding, and the results are included together with additional experiments in Appendix F. For domain adversarial training, we use the standard procedure (Ganin et al., 2016), with an initial $\mu = 0$ that follows an exponentially increasing schedule per epoch of training for the range $[0, 1]$. In addition, we assess the impact of the adversarial training procedure in TE-CDE by training a version of our method *without* domain adversarial training, i.e. constant $\mu = 0$ in the loss function (Eq. 9).

Experimental details. Implementation details, including hyper-parameters, can be found in Appendix E. Unless otherwise stated, each experiment is run with 10,000 patients for training, 1,000 for validation and 10,000 for testing.

5.2. Impact of time-dependent confounding across varying sampling intensities

A key challenge when learning from longitudinal data is accounting for bias introduced by time-dependent confounding. Therefore, it is essential to assess the impact of

---

[4]https://github.com/seedatnabeel/TE-CDE
[5]https://github.com/vanderschaarlab/mlforhealthlabpub/tree/main/alg/TE-CDE

time-dependent confounding on counterfactual estimation. We measure performance via normalized RMSE, where the RMSE is normalized by the maximum tumor volume $V_{\max} = 1150\text{cm}^3$. As discussed, time-dependent confounding is controlled by parameters γ_c, γ_r in the treatment assignment policies. We evaluate the benchmarks under increasing degrees of time-dependent confounding by setting $\gamma_c = \gamma_r = \gamma = \{2, 4, 6, 8, 10\}$.

As previously discussed, a patient's condition often affects the frequency of observation. The state-based variability in sampling intensity based on the clinical state is controlled by κ in the simulation environment. We repeat the experiment for multiple values of the scaling factor $\kappa = \{1, 5, 10\}$.

Figure 3 shows the results for counterfactual estimation for different levels of sampling intensity κ. As expected, the performance of all models degrades with increasing time-dependent confounding. However, TE-CDE achieves the lowest counterfactual estimation RMSE for all values of sampling intensity κ and across all values of time-dependent confounding γ. The divergence is most pronounced for increasing γ, with $\gamma = 10$ leading to a 36% decrease in RMSE for TE-CDE compared to CRN, the next best performing method. The superior performance of TE-CDE in all settings highlights the benefit of the continuous-time approach adopted compared to RNN-based approaches.

Comparing the RNN-based models, CRN outperforms RMSN, matching the conclusions in the regularly sampled setting reported in Bica et al. (2020b). Overall, however, the results highlight the limitation of RNN-based models in the irregularly sampled setting.

Finally, we characterize the value of domain adversarial training in TE-CDE by comparing it to the case when $\mu = 0$ (i.e. no domain adversarial training). TE-CDE ($\mu = 0$) suffers significant performance degradation with a higher RMSE in all scenarios that grows as the degree of time-dependent confounding increases. This clearly demonstrates the practical benefit of the adversarial training approach to learning balancing representations.

5.3. Treatment-conditioned sampling

In the previous section, we consider the realistic scenario where the severity of the patients' condition governs the intensity of the observation process, i.e. sicker patients with higher stages are observed more frequently. In addition, the treatment regimen itself will often also influence the observation pattern of the patient, i.e. patients undergoing different treatments are monitored differently.

To simulate this phenomenon, we adjust the base sampling intensity κ depending on whether the patient is treated or untreated, thereby altering the state-dependent sampling variability between the two treatment groups. We set $\kappa = 10$

for treated patients and $\kappa = 1$ for untreated patients. We fix the time-dependent confounding as $\gamma = 4$.

Consistent with the previous experiment, TE-CDE significantly outperforms the benchmark models (Table 2). There is a divergence in performance between the treated and untreated populations. This is largely explained by differences in the severity of the clinical condition between the two populations: the majority of untreated patients (77%) remain in cancer stage S1A (i.e. $\max s_t = 0$). Due to the reduced state transition, these patients naturally have lower variability in tumor volume over the trajectory, leading to lower error for all methods. Overall, TE-CDE outperforms all methods globally and for treated and untreated patients.

Table 2: Normalized RMSE (%) for counterfactual estimation with κ conditioned on treatment, where $\kappa_{treated} = 10$ and $\kappa_{untreated} = 1$.

Model	Overall	Treated	Untreated
TE-CDE	1.18 ± 0.05	1.56 ± 0.06	0.22 ± 0.02
CRN	1.57 ± 0.06	1.97 ± 0.05	0.64 ± 0.08
RMSN	3.06 ± 0.09	3.12 ± 0.07	2.83 ± 0.08

5.4. Forecasting at additional time horizons

We have assessed the ability to estimate counterfactual outcomes at the subsequent observation time determined by the Hawkes process described in Section 5.1. To further validate our method, we assess counterfactual estimation at subsequent observation times that are further in the future.

As an illustrative example, we evaluate counterfactual estimation at time t_{k+n}, i.e. estimate tumor volume $y_{t_{k+n}}$. Similar to other experiments, we vary the degree of time-dependent confounding $\gamma = \{2, 4, 6, 8, 10\}$, fix $\kappa = 10$, and set the forecasting horizon $n = 5$ (see Appendix F.8 for other time horizons). As expected, it is more challenging to estimate counterfactuals further in the future. As shown in Figure 4, similar trends are observed as the setting of Section 5.2 (Figure 3). TE-CDE outperforms both CRN and RMSN for all γ, with a greater performance differential as the degree time-dependent confounding increases. For $\gamma = 10$, there is a 40% reduction in RMSE for TE-CDE.

5.5. Treatment selection

To demonstrate how models such as TE-CDE could be used in decision support and the potential impact assisting clinical decision makers, we must assess performance in ways beyond counterfactual estimation. One such clinically relevant evaluation is whether the best treatment was selected. This is important since reduced error in counterfactual es-
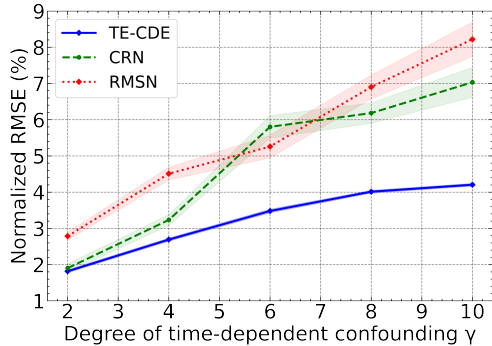
Figure 4: Counterfactual estimation at time t_{k+n} ($n = 5$) for varying levels of time-dependent confounding γ.

timation does not necessarily result in improved clinical outcomes.

We define the "correct" treatment selection as the treatment that minimizes the tumor volume at time t_{k+n} (i.e. $y_{t_{k+n}}$).

We adopt the same experimental setup as Section 5.4 and set the forecasting horizon $n = 5$, sampling intensity $\kappa = 10$, and vary time-dependent confounding $\gamma = \{2, 4, 6, 8, 10\}$.

Figure 5 shows decreasing accuracy for all methods as time-dependent confounding increases. However, for all values of γ, TE-CDE outperforms both CRN and RMSN, more frequently selecting the optimal treatment. Similar to the other experiments, as γ increases, the performance gap between TE-CDE and both CRN and RMSN increases, with c. 4% difference in absolute treatment selection accuracy at $\gamma = 4$ increasing to c. 10% at $\gamma = 10$. This experiment emphasizes that differences in counterfactual estimation result in meaningful differences in treatment selection accuracy.
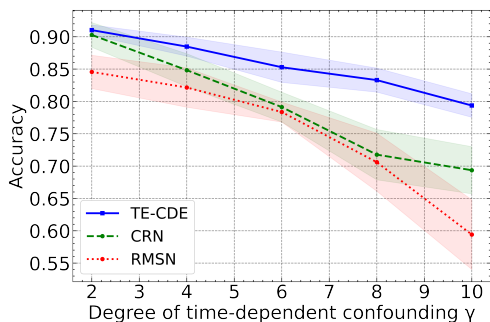

Figure 5: Treatment accuracy under varying degrees of time-dependent confounding γ.

5.6. Additional experiments

We perform a number of additional experiments to further validate TE-CDE. In Appendix F.1, we show how uncertainty estimates can be obtained from TE-CDE and then

used to rank counterfactual estimates such that uncertain samples can be deferred to clinicians to improve outcomes. Appendix F.2 compares the data efficiency of the different methods, which is useful in clinical settings with limited labeled data. In Appendix F.3, we explore the latent representation of TE-CDE over time to: (1) highlight that the latent states z_t learned by TE-CDE indeed are treatment-invariant representations and (2) investigate clinical insights that can be ascertained from the latent representations.

6. Conclusion

State-of-the-art methods for counterfactual estimation are predicated on the assumption of regular and evenly spaced data sampling. However, real-world clinical time series are often irregular. To address this challenge, we introduce TE-CDE, a model that learns to perform counterfactual estimation in continuous time from irregularly sampled observational data with time-dependent confounding. Additionally, we propose a controlled simulation environment for medically realistic irregularly sampled time series. In experiments in a variety of irregular settings, we demonstrate that TE-CDE provides improvements over current state-of-the-art methods.

Counterfactual estimation has the potential to assist clinicians with "what-if" decision-making. However, when deploying such models in healthcare settings, there are risks, e.g. inaccurate predictions. Furthermore, trade-offs between outcomes and possible treatment side effects are not accounted for by such models. To mitigate possible adverse effects, counterfactual estimates should be part of a "human-in-the-loop" paradigm, allowing experts to complement predictions with domain knowledge to improve patient outcomes.

We also note that while some aspects of irregularly sampled data are naturally addressed through our formulation, this work is simply a step in the right direction and our proposed solution only partially addresses the complexities of irregular sampling. In particular more is needed to address a number of aspects, such as informative sampling, that are widely prevalent in healthcare. We hope that this serves as a motivation for future work.

Acknowledgements

References

Alaa, A. M. and van der Schaar, M. Bayesian inference of individualized treatment effects using multi-task gaussian processes. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 3427–3435, 2017.

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.

Bao, Y., Kuang, Z., Peissig, P., Page, D., and Willett, R. Hawkes process modeling of adverse drug reactions with longitudinal observational data. In *Machine learning for healthcare conference*, pp. 177–190. PMLR, 2017.

Bartsch, H., Dally, H., Popanda, O., Risch, A., and Schmezer, P. Genetic risk profiles for cancer susceptibility and therapy response. *Cancer Prevention*, pp. 19–36, 2007.

Bellot, A. and van der Schaar, M. Policy analysis using synthetic controls in continuous-time. *In Proceedings of 38th International Conference on Machine Learning (ICML 2021)*, 2021.

Bica, I., Alaa, A., and Van Der Schaar, M. Time series deconfounder: Estimating treatment effects over time in the presence of hidden confounders. In *Proceedings of the 37th International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2020a.

Bica, I., Alaa, A. M., Jordon, J., and van der Schaar, M. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. *In Proceedings of 8th International Conference on Learning Representations (ICLR 2020)*, 2020b.

Bica, I., Alaa, A. M., Lambert, C., and Van Der Schaar, M. From real-world patient data to individualized treatment effects using machine learning: current and future methods to address underlying challenges. *Clinical Pharmacology & Therapeutics*, 109(1):87–100, 2021.

Chen, R. T., Rubanova, Y., Bettencourt, J., and Duvenaud, D. Neural ordinary differential equations. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 6572–6583, 2018.

De Brouwer, E., Simm, J., Arany, A., and Moreau, Y. Gru-ode-bayes: Continuous modeling of sporadically-observed time series. In *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, 2019.

Gal, Y. and Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

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.

Gershenwald, J. E., Scolyer, R. A., Hess, K. R., Sondak, V. K., Long, G. V., Ross, M. I., Lazar, A. J., Faries, M. B., Kirkwood, J. M., McArthur, G. A., et al. Melanoma staging: evidence-based changes in the american joint committee on cancer eighth edition cancer staging manual. *CA: a cancer journal for clinicians*, 67(6):472–492, 2017.

GPy. GPy: A gaussian process framework in python. `http://github.com/SheffieldML/GPy`, since 2012.

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.

Hawkes, A. G. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90, 04 1971. doi: 10.1093/biomet/58.1.83. URL `https://doi.org/10.1093/biomet/58.1.83`.

Hawkes, A. G. and Oakes, D. A cluster process representation of a self-exciting process. *Journal of Applied Probability*, 11(3):493–503, 1974.

Hernán, M. Á., Brumback, B., and Robins, J. M. Marginal structural models to estimate the causal effect of zidovudine on the survival of hiv-positive men. *Epidemiology*, pp. 561–570, 2000.

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

Ilg, E., Cicek, O., Galesso, S., Klein, A., Makansi, O., Hutter, F., and Brox, T. Uncertainty estimates and multi-hypotheses networks for optical flow. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 652–667, 2018.

Jesson, A., Mindermann, S., Shalit, U., and Gal, Y. Identifying causal-effect inference failure with uncertainty-aware models. *Advances in Neural Information Processing Systems*, 33, 2020.

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

Johansson, F. D., Sontag, D., and Ranganath, R. Support and invertibility in domain-invariant representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 527–536. PMLR, 2019.

Johansson, F. D., Shalit, U., Kallus, N., and Sontag, D. Generalization bounds and representation learning for estimation of potential outcomes and causal effects. *arXiv preprint arXiv:2001.07426*, 2020.

Kidger, P., Morrill, J., Foster, J., and Lyons, T. Neural controlled differential equations for irregular time series. *34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, 2020.

Lee, Y., Lim, K. W., and Ong, C. S. Hawkes processes with stochastic excitations. In *International Conference on Machine Learning*, pp. 79–88. PMLR, 2016.

Li, S. and Fu, Y. Matching on balanced nonlinear representations for treatment effects estimation. In *NIPS*, 2017.

Lim, B., Alaa, A. M., and van der Schaar, M. Forecasting treatment responses over time using recurrent marginal structural networks. *NeurIPS*, 18:7483–7493, 2018.

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

Lyons, T. J., Caruana, M., and Lévy, T. *Differential equations driven by rough paths*. Springer, 2007.

Mansournia, M. A., Danaei, G., Forouzanfar, M. H., Mahmoodi, M., Jamali, M., Mansournia, N., and Mohammad, K. Effect of physical activity on functional performance and knee pain in patients with osteoarthritis: analysis with marginal structural models. *Epidemiology*, pp. 631–640, 2012.

Mansournia, M. A., Etminan, M., Danaei, G., Kaufman, J. S., and Collins, G. Handling time varying confounding in observational research. *bmj*, 359, 2017.

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

Platt, R. W., Schisterman, E. F., and Cole, S. R. Time-modified confounding. *American journal of epidemiology*, 170(6):687–694, 2009.

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., Hernán, M., and Brumback, B. Marginal structural models and causal inference in epidemiology. *Epidemiology (Cambridge, Mass.)*, 11(5):550–560, 2000.

Robins, J. M. Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics-Theory and methods*, 23(8):2379–2412, 1994.

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

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

Roueff, F., Von Sachs, R., and Sansonnet, L. Locally stationary hawkes processes. *Stochastic Processes and their Applications*, 126(6):1710–1743, 2016.

Rubanova, Y., Chen, R. T., and Duvenaud, D. Latent ODEs for irregularly-sampled time series. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pp. 5320–5330, 2019.

Ryalen, P. C., Stensrud, M. J., and Røysland, K. The additive hazard estimator is consistent for continuous-time marginal structural models. *Lifetime data analysis*, 25(4): 611–638, 2019.

Saarela, O. and Liu, Z. A flexible parametric approach for estimating continuous-time inverse probability of treatment and censoring weights. *Statistics in medicine*, 35 (23):4238–4251, 2016.

Schisterman, E. F., Cole, S. R., and Platt, R. W. Overadjustment bias and unnecessary adjustment in epidemiologic studies. *Epidemiology (Cambridge, Mass.)*, 20(4):488, 2009.

Schulam, P. and Saria, S. Reliable decision support using counterfactual models. *Advances in Neural Information Processing Systems*, 30:1697–1708, 2017.

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.

Tatekawa, K., Iwata, H., Kawaguchi, T., Ishikura, S., Baba, F., Otsuka, S., Miyakawa, A., Iwana, M., and Shibamoto, Y. Changes in volume of stage i non-small-cell lung cancer during stereotactic body radiotherapy. *Radiation oncology*, 9(1):1–5, 2014.

Van der Maaten, L. and Hinton, G. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

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

Wang, Y. and Blei, D. M. The blessings of multiple causes. *Journal of the American Statistical Association*, 114(528): 1574–1596, 2019.

Yao, L., Li, S., Li, Y., Huai, M., Gao, J., and Zhang, A. Representation learning for treatment effect estimation from observational data. *Advances in Neural Information Processing Systems*, 31, 2018.

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

Zhang, H., Gao, X., Unterman, J., and Arodz, T. Approximation capabilities of neural ODEs and invertible residual networks. In *International Conference on Machine Learning*, pp. 11086–11095. PMLR, 2020.

A. Notation Summary

Table 3: Summary of notation used in the paper

Notation	Explanation
$\mathbf{X}$	Covariate process
A	Treatment process
Y	Outcome path
$Y(A = a)$	Potential outcome of Y
t	time t
N	Counting Process
$\mathcal{F}_t$	Filtration
λ	Intensity process
$\mathbf{Z}$	Latent path
f_θ	Encoder latent vector field
f_ϕ	Decoder latent vector field
$\mathbf{X}_t$	Covariates at time t
A_t	Treatment at time t
Y_t	Outcome at time t
h_v	Outcome prediction network
h_a	Treatment prediction network
μ	Parameter controlling the trade-off between treatment and outcome prediction
$\mathcal{L}^{(a)}$	Outcome loss
$\mathcal{L}^{(a)}$	Treatment loss

Notation for the tumor growth dynamics can be found in Appendix C.1

B. Extended Related Work

Treatment effect estimation with static data.
A large body of work has proposed causal inference methods applied to observational data in the static setting. The methods can be grouped as: (a) Representation learning (Johansson et al., 2016; Shalit et al., 2017; Yoon et al., 2018; Li & Fu, 2017; Yao et al., 2018), (b) Matching and Re-weighting (Rosenbaum & Rubin, 1983) and (c) Tree-based (Hill, 2011; Wager & Athey, 2018).

While the aforementioned methods showed good performance, the static setting has two properties that simplify the estimation, namely (1) the treatment is allocated only once and does not change over time, and (2) the confounders and the outcomes are static variables rather than time series. Additionally, the static setting does not naturally apply to the medical studies based on longitudinal data (e.g. EHRs), where the treatments, confounders and outcomes are all time-varying (Hernán et al., 2000; Schisterman et al., 2009; Mansournia et al., 2012). The wealth of information being routinely collected as a part of the electronic health record (EHR) provides an unprecedented opportunity to discover appropriate clinical recommendations for patients given

Treatment effect estimation with longitudinal, time-varying data.
A number of methods have been proposed for estimating the effect of time-varying exposures in the presence of time-dependent confounding, including g-computation (Robins, 1986), Structural Nested Models (Robins, 1994), and Marginal Structural Models (MSMs) (Robins et al., 2000). However, MSMs are sensitive to model mis-specification, when computing the propensity weights and hence when estimating outcomes.

Methods for the longitudinal setting with temporal confounding such as Marginal Structural Models (MSMs) (Robins et al., 2000), have thus been extended to address some of these issues. The Recurrent Marginal Structural Model (RMSN) (Lim et al., 2018) aimed to reduce variance in the weights using recurrent neural networks to estimate the propensity weights and to build the outcome model.

Recently, the Counterfactual Recurrent Network (CRN) (Bica et al., 2020b) combined a RNN architecture with adversarial training to balance the covariate distributions in different treatment regimes. That said, the aforementioned methods applied to longitudinal, time-varying data make unrealistic assumptions that the measurements are made at regular intervals, fully aligned across all patients, or do not contain missing values.

The Counterfactual Gaussian Process (CGP) (Schulam & Saria, 2017) made no such assumptions on data sampling using a generative model to jointly model the outcome and the treatment allocation. However, CGP makes strong assumptions about model structure by using Bayesian non-parametric methods, which makes them unsuitable to handle heterogeneous treatment effects arising from baseline variables or settings with multiple treatment outcomes. In addition, CGP does not directly address the issue of time-dependent confounding.

While in this work we assume no hidden confounding, we do note a body of work such as Wang & Blei (2019) and Bica et al. (2020a) exists which attempts to adjust for hidden confounders in observational data.

Neural differential equations.
We consider the trajectory a patient under-going treatment as a dynamical system. Neural ordinary differential equations (Neural ODEs) and extensions have been shown to successfully model the continuous-time evolution of dynamical systems with differential equations learned by neural networks (Chen et al., 2018; Rubanova et al., 2019; De Brouwer et al., 2019). However, a key limitation of Neural ODE approaches is that the entire trajectory is determined by the initial condition at $t = 0$ and the equation does not account for any information available at $t > 0$.

Recent works on neural controlled differential equations (Neural CDE) (Kidger et al., 2020; Morrill et al., 2021) address this shortcoming by allowing incoming information to modulate the dynamics. This ability is naturally useful in a clinical setting, as not only can we model the continuous-time latent state evolution of a patient trajectory, but also we account for incoming data (e.g. treatment changes) that modulate the dynamics of system.

As was discussed in the main paper, (Gwak et al., 2020) and (Bellot & van der Schaar, 2021) have used neural differential equations for intervention modeling. While somewhat related by virtue of using neural differential equations, these works are not applicable to our setting.

(Gwak et al., 2020) used two separate neural ODEs to model the intervention and outcome processes. That said, the method is not applicable to treatment effect estimation in clinical settings as it does not integrate time-varying covariates or adjust for confounding.

On the other hand (Bellot & van der Schaar, 2021) proposed to model treatment effects in continuous-time in the context of synthetic controls. The setting however, is completely different as it only considered a single intervention and the synthetic control approach is not applicable more generally.

As a consequence, to the best of our knowledge TE-CDE is the first method for counterfactual estimation and treatment effects to leverage the mathematics of neural differential equations and more specifically neural controlled differential equations .

C. Simulation Environment

C.1. Pharmacokinetic-Pharmacodynamic Model

Pharmacokinetic-Pharmacodynamic models are bio-mathematical models which represent dose-response relationships. This enables clinicians to understand possible patient response and hence propose optimal treatments.

In our setting we wish to estimate counterfactual outcomes. Since counterfactuals do not exist in practice, we make use of a synthetic data generated via a PK-PD model. In particular, we use a state-of-the-art bio-mathematical by (Geng et al., 2017).

The PK-PD model is used to model lung cancer tumor growth under effects of chemotherapy and radiotherapy. For evaluation of counterfactuals such a model allows us to generate data of patient tumor volume outcomes under different possible treatments.

Model of tumor growth
The tumor volume at time t after diagnosis is modeled as follows:

$$\frac{dV(t)}{dt} = \left(\underbrace{\rho \log \left(\frac{K}{V(t)} \right)}_{\text{Tumor growth}} - \underbrace{\beta_c C(t)}_{\text{Chemotherapy}} - \underbrace{\left(\alpha_r c(t) + \beta_r c(t)^2 \right)}_{\text{Radiotherapy}} + \underbrace{e_t}_{\text{Noise}} \right) V(t), \tag{12}$$

The parameters $K, \rho, \beta_c, \alpha_r, \beta_r$ for each simulated patient are sampled from the prior distributions described in (Geng et al., 2017) and details are outlined in Table 3 below. We include $e_t \sim \mathcal{N}(0, 0.01^2)$ as random noise to account for randomness in the tumor growth.

Table 4: Outline of different parameters used in the PK-PD Model

Model	Variable	Parameter	Distribution	Parameter Value $(\mu, \sigma))$
Tumor growth	Growth parameter	ρ	Normal	$7.00 \times 10^{-5}, 7.23 \times 10^{-3}$
	Carrying capacity	K	Constant	30
Radiotherapy	Radio cell kill (α)	α_r	Normal	0.0398, 0.168
	Radio cell kill (β)	β_r	-	Set s.t. α/β=10
Chemotherapy	Chemo cell kill	β_c	Normal	0.028, 0.0007

Heterogeneous responses. In the simulation we wish to incorporate heterogeneity among patient responses to match the real-world where individualized treatment effect can vary due to factors such as gender or genetics (Bartsch et al., 2007).

Similar to (Bica et al., 2020b) and (Lim et al., 2018), the means are adjusted for β_c and α_r by creating three groups of patients (i.e. to represent three types of patients with heterogeneity in treatment response).

For patient group 1, we update the mean of α_r such that $\mu(\alpha_r) = 1.1 \times \mu(\alpha_r)$ and for patient group 3, we update the mean of α_c such that $\mu(\alpha_c) = 1.1 \times \mu(\alpha_c)$.

Drug concentrations ($C(t)$ and $d(t)$).
The chemotherapy drug concentration $C(t)$ follows an exponential decay relationship with a half life of one day:

$$C(t) = \tilde{C}(t) + C(t-1)/2, \tag{13}$$

where $\tilde{C}(t)$ represents a 5.0mg/m^3 dosage of Vinblastine given at time t.

The radiotherapy concentration $d(t)$ represents $2.0Gy$ fractions of radiotherapy given at timestep t, where Gy is the Gray ionizing radiation dose.

Time-dependent confounding.

We incorporate time-dependent confounding into the data generating process by modeling chemotherapy and radiotherapy assignment as Bernoulli random variables, with probabilities p_c and p_r which depend on tumor diameter:

$$p_c(t) = \sigma\left(\frac{\gamma_c}{D_{\max}}(\bar{D}(t) - \delta_c)\right) \qquad\qquad p_r(t) = \sigma\left(\frac{\gamma_r}{D_{\max}}(\bar{D}(t) - \delta_r)\right), \qquad (14)$$

where $D_{\max} = 13\text{cm}$ is the maximum tumor diameter, $\theta_c = \theta_r = D_{\max}/2$ and $\bar{D}(t)$ is the average tumor diameter. The degree of time-dependent confounding is controlled by γ_c and γ_r, where increasing $\gamma_{\{c,r\}}$ increases the amount of the time-dependent confounding.

C.2. Cancer Staging

As described previously, patient outcomes can evolve through different states of disease over the course of their observational trajectory. These different states correspond to clinical guidelines used to grade cancer stages. In clinical settings, patient monitoring or observation frequency often varies for different clinical stages.

For example, patients in more severe states with greater tumor volume are monitored more frequently and hence have a greater degree of sampling. However, as the patient's state/condition improves the measurements reduce in frequency (i.e. routine and less frequent follow-ups).

Thus, in the simulated data, the observational sampling rates closely follow this pattern to reflect healthcare data, shown in Figures 6 and 7. This induces sampling irregularity both within a trajectory as the patient state changes over time as well as irregularity between different patient trajectories.

This irregularity is modeled via a Hawkes process as discussed in the main paper. Note, however, that $\alpha/\beta < 1$ in the Hawkes process, thereby ensuring local stationarity of the Hawkes process (Roueff et al., 2016).

We categorize tumor stages as per the American Joint Committee on Cancer (AJCC) Tumor, Node, Metastasis (TNM) staging system (Gershenwald et al., 2017):

- Stage 0: No cancer

- Stage 1A: 0cm < tumor ≤ 3cm

- Stage 1B: 3cm < tumor ≤ 4cm

- Stage 2: 4cm < tumor ≤ 5cm

- Stage 3 & 4 [6]. : tumor > 5cm

The observational outcome is the volume of the lung cancer tumor, however the classification of stage is based on tumor diameter. Hence, we assume that the tumor has a perfectly spherical tumor shape similar to (Tatekawa et al., 2014) and hence can compute the diameter as per Equation 15.

$$volume = \pi/6 \times (diameter)^3 \qquad (15)$$

---

[6]We group these stages as stage 4 simply refers to the tumor invading a specific structure or anatomical space
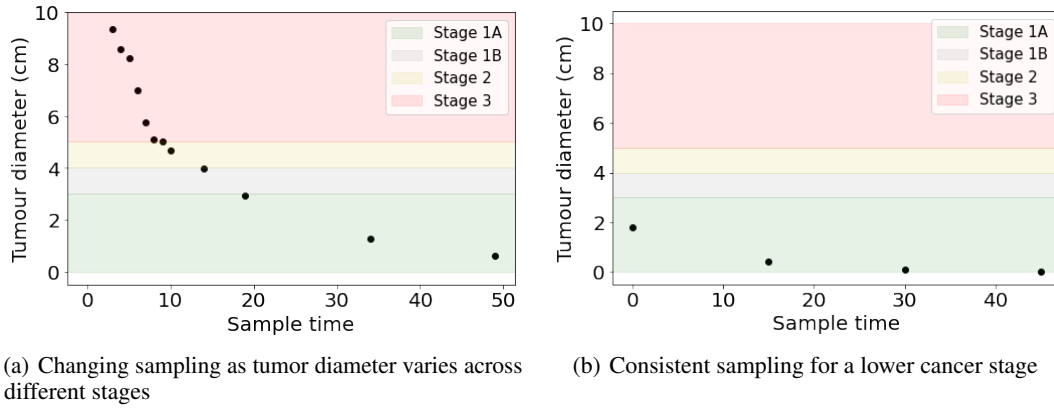
(a) Changing sampling as tumor diameter varies across different stages

(b) Consistent sampling for a lower cancer stage

Figure 6: Illustrations of different potential sampling patterns

(a) Oscillation in tumor diameter and stage for more severe disease

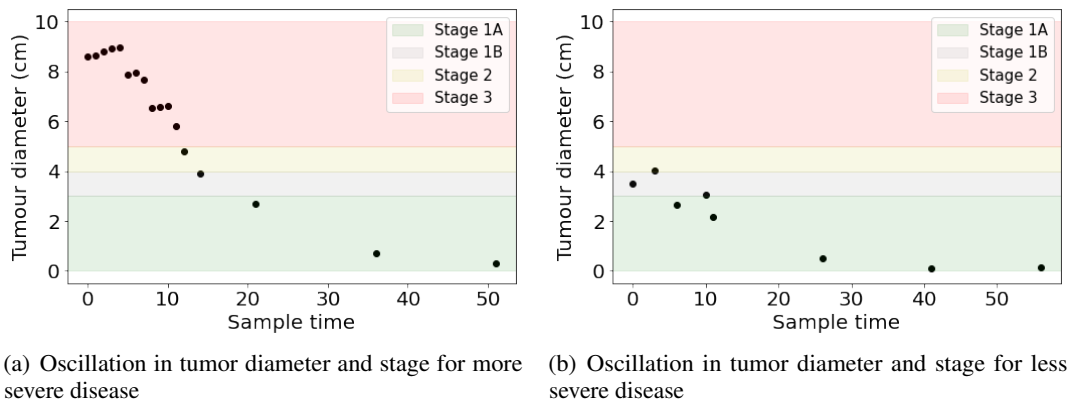(b) Oscillation in tumor diameter and stage for less severe disease

Figure 7: Illustration that tumor diameter can oscillate and even increase over a trajectory - for example due to non-response to treatment.

D. TE-CDE

In this section, we highlight some theoretical nuances and explanations of TE-CDE not captured in the main paper, as well as outlining further implementation details.

D.1. Additional Methodological Details

TE-CDE latent path construction:
In order to construct the neural controlled differential equation a control path X_t must be defined. The vector field trajectory is driven by (i.e. controlled) by the time-varying X_t. Recall we have an observational time series of multi-dimensional data points $((x_{t_0}, a_{t_0}, y_{t_0}), (x_{t_1}, a_{t_1}, y_{t_1}), ...(x_{t_k}, a_{t_k}, y_{t_k}))$. The control path X_t is then a continuous approximation of these values, which can be obtained through interpolation (choice of interpolation is discussed later).

The latent state z_t is then the solution of a CDE, controlled by X_t described formally by Equation 5 and 6 below.

$$\mathbf{Z}_{t_0} = g_\eta(\mathbf{X}_{t_0}, A_{t_0}, Y_{t_0}), \tag{16}$$

$$\mathbf{Z}_t = \mathbf{Z}_{t_0} + \int_{t_0}^{t} f_\theta(\mathbf{Z}_s) \frac{d[\mathbf{X}_s, A_s, Y_s]}{ds} ds, \tag{17}$$

where the integral is a Riemann-Stieltjes integral.

Differences between Neural CDEs and Neural ODEs:
The primary difference between a Neural CDE and a Neural ODE lies with the control path X_t. The control path addresses a shortcoming of a neural ODE, where the latent state z_t in an ODE is decided by z_{t0}, hence is unsuitable to adapt to incoming data due to reliance on the initial value. Instead, neural CDEs have a latent state z_t controlled by another time series X_t. That said, Neural CDEs can be thought of as a generalization of a neural ODE, as neural CDEs can reduce to neural ODEs for the specific case where $X_t = t$.

Training efficiency:
TE-CDE and neural CDEs in general make use of the adjoint backpropagation method. This means that the memory requirements when training TE-CDE is $\mathcal{O}(T + V)$ where T=$t_1 - t_0$ (time spacing) and V is the size of the vector field.

Selecting an interpolation method:
Since we are performing counterfactual estimation, we thus need to retain a causal interpretation from TE-CDE. When constructing the continuous control path X_t, we need to be careful as all interpolation methods except Linear and Rectilinear are non-causal in nature. i.e. other interpolation methods need future points in time to approximate the signal. This means that when using Neural CDEs in scenarios that require causal interpretations, such as TE-CDE, linear or rectilinear interpolation must be used.

TE-CDE solves a well-posed problem:
We highlight that TE-CDE solves a well-posed problem. It was shown by (Lyons et al., 2007) that CDEs solve a well-posed problem under the condition of Lipschitz continuity. We aim to show that TE-CDE still fits the well-posed paradigm. In our architecture we make use of ReLU and softmax activations as well as common neural network layers such dropout which are Lipschitz constant. Thus, the Lipschitz conditions still hold which means that TE-CDE fulfills the following conditions of a well-posed problem: (1) a solution exists, (2) the solution is unique, (3) the behaviour of the solution changes continuously based on the input data. By virtue of TE-CDE solving a well-posed problem, we note that the training of TE-CDE is thus theoretically stable.

Representing data continuously:
TE-CDE requires a continuous-time embedding of the observational data. Whilst this has numerous benefits, it also introduces challenges. The benefit is firstly it allows us to naturally handle irregularly sampled data and secondly, TE-CDE produces a solution (i.e. the latent state z_t) defined at all points in time (i.e. continuously).

The challenge, however, is that there may be discontinuities in $A_{\leq t}$ (e.g. treatment is applied in discrete stages). This results in a piecewise smooth control path (not smooth due to linear interpolation). Thus, in our case when using an adaptive solver with a piecewise smooth control path, we then need to explicitly inform the solver about the jumps between pieces so that its integration steps may align with them. This can be achieved using the `jump_t` argument in the CDE solver (Chen et al., 2018; Kidger et al., 2020).

If this is not done the adjoint backpropagation method with an adaptive ODE solver can be numerically and computationally expensive (i.e. slower), as the solver must then locate the discontinuities and then slow down to resolve them.

D.2. TE-CDE Implementation Details

We discuss the implementation details for TE-CDE in terms of encoder-decoder architecture, optimization and general comments about training.

Encoder-Decoder Architecture:
The encoder and decoder of TE-CDE both use Neural CDEs with the same architectures - the difference being the observations for the encoder and decoder (i.e. different inputs). The integrand of the neural CDE (i.e. f) is a 2-layer neural network with hidden states of size=128. The dimensionality of the latent state z is 8. We use linear interpolation when defining the control path X_t.

Optimization:
We make use of Stochastic Gradient Descent (SGD) as implemented in the Pytorch framework. The ODE solver used to extrapolate the latent state z_t is an adaptive step size solver namely the Dormand–Prince (`dopri5`) method, which is a member of the Runga-Kutta family of ODE Solvers. We make use of the `torchcde` package with `torchdiffeq` backend. As discussed to account for the discontinuities and piecewise smooth control path we make use of the `jump_t` argument.

General comments:
Both encoder and decoder are trained for 100 epochs each. That said we also include early stopping in the training protocol based on the validation loss, with patience=5. When MC Dropout is included, we use a dropout probability=0.1. The model was implemented with Pytorch and TorchCDE and was trained and evaluated on a single Nvidia P100 or T4 GPU. We did not explicitly tune hyperparameters (activation function, ODE Solver) for performance. Further tuning could be done for these aspects based on a validation set. However, we did tune the learning rate (lr = 1e-3, **1e-4***, 1e-5, 1e-6) based on performance on the validation set.

E. Implementation details: Benchmark algorithms

Due to the underlying usage of Recurrent Neural Networks (RNNs) both CRN and RMSN rely on the assumption of regularly sampled data. Thus, for the irregularly sampled data setting, which effectively means un-observed data samples for different discrete time steps, as is commonly done we include a pre-processing step for RNN-based models, where we divide the timeline equally, interpolate and impute the "un-observed" observations.

E.1. Counterfactual Recurrent Network (CRN)

We benchmark CRN using the Tensorflow implementation of (Bica et al., 2020b) as per [7]. We pre-process the data for the irregularly sampled setting as described.

E.2. Recurrent Marginal Structural Network (RMSN)

We benchmark RMSN using the Tensorflow implementation of (Lim et al., 2018) as per [8]. We pre-process the data for the irregularly sampled setting as described. We tune parameters as described by the authors in their repo.

E.3. Gaussian Process (GP)

The Gaussian Process (GP) model is evaluated as a continuous-time comparison and is inspired by (Schulam & Saria, 2017). The method however, like prior works using GPs including (Schulam & Saria, 2017) does not account for time-dependent confounding. Similar to (Schulam & Saria, 2017), we use a Matérn 3/2 kernel with variance=0.22 and lengthscale=8.0. The benchmark algorithm is implemented using GPy (GPy, since 2012).

---

[7]`https://github.com/ioanabica/Counterfactual-Recurrent-Network`
[8]`https://github.com/sjblim/rmsn_nips_2018`

F. Additional Experiments

We perform a number of additional experiments to further validate TE-CDE.

F.1. Uncertainty Based Exclusion of Estimates

Uncertainty concerning the counterfactual estimates is an important problem to consider as estimation errors can lead to harmful decisions being taken based on the incorrect estimates. Whilst (Jesson et al., 2020) have examined uncertainty in the static setting, this area is under-studied in longitudinal setting. We aim to illustrate how uncertainty can be modeled within the TE-CDE framework. We then evaluate the utility of the uncertainty estimates to (a) improve model outcomes and counterfactual estimates and (b) examine how uncertainty could be utilized in clinical workflows such that uncertain cases can be flagged and deferred to a clinician.

Bayesian methods such as variational inference or MCMC can be computationally intensive. Hence, we propose to augment and re-train our TE-CDE model with MC-Dropout (Gal & Ghahramani, 2016). This then allows us to sample N counterfactual outputs $y^{(i)}$, where each dropout mask corresponds to a sample from the approximate posterior distribution $q(\theta)$. The variance of these samples is then representative of the epistemic uncertainty (model uncertainty) in counterfactual estimation.

$$\tilde{\mathbf{y}} = \frac{1}{N} \sum y^{(i)} \qquad\qquad \sigma^2 = \text{Epistemic uncertainty} = var(\mathbf{y}) \qquad (18)$$

We highlight the utility of the uncertainty estimates of TE-CDE to improve model outcomes and counterfactual estimates, by deferring uncertain samples. As discussed, the uncertainty is quantified by the variance of the MC-Sampled counterfactual estimates (epistemic uncertainty). The uncertainty over the trajectory at every timestep is then averaged with the uncertainty estimates represented as $u^{all} = (u_1^a, u_2^a, u_3^a...u_n^a)$. The u^{all} across all patients is then rank sorted based on the uncertainty metric.

As an illustrative example, we evaluate the quality of TE-CDE's uncertainty estimates by modeling how the RMSE varies as a function of the proportion of "uncertain" data excluded. We posit that a useful measure of uncertainty should lead to a lower RMSE on the retained subset with highest confidence. Hence, we define a meaningful representation of uncertainty as one where the error curve decreases for increasing exclusion of uncertain data. Such a measure could then be used to improve clinical workflows by deferring uncertain estimates to clinicians.

The aforementioned test ascertains if the uncertainty estimate can correctly order the counterfactual estimates. However, a limitation as a result of MC Dropout is the coupling with the model which introduces an implicit dependency on the model that generates the prediction's and uncertainty estimates. Hence, we cannot directly compare different approaches.

To address the coupling we utilize the concept of Sparsification from (Ilg et al., 2018). This is a novel application from optical flow to causal inference. We define an Oracle ($u^{oracle} = (u_1^o, u_2^o, u_3^o...u_n^o)$). The oracle has the best possible ordering of samples, with bounds based on knowing the ground truth and thereby ordering by magnitude of the true error. Thus, the oracle is the lower bound on performance. We can then compute the difference between uncertainty based exclusion and the oracle exclusion given by Equation 8. The oracle is model dependant and by computing the difference we marginalize out the impact of the oracle. Hence, we can then use the sparsification error to directly compare different uncertainty estimates.

$$\text{Sparsification error} = \sum_{i=1}^{n} u_i^a - u_i^o \qquad (19)$$

In order, to produce a single comparative metric to compare models, the area under the sparsification error curve (AUSE) can be used to compare the different models and uncertainty measures. Smaller values correspond to better performance as the measure would then be closer to the oracle lower bound (i.e. closer to ground truth).

Baseline: We use two baselines to validate the utility of TE-CDE's uncertainty estimates: (1) Oracle: best possible exclusion and (2) Random: randomly exclude a proportion of samples. Random exclusion should not result in a reduction in RMSE.

Results: The exclusion curves are illustrated in Figure 8 (a) demonstrates a decreasing curve for sorting by variance/epistemic uncertainty. This suggests TE-CDE can be used to produce a useful ranking by uncertainty and performs better than the random baseline. This is corroborated by the Sparsification curve and AUSE (Figure 8 (b)).

When comparing exclusion methods it highlights that ordering by the variance/epistemic uncertainty significantly outperforms random exclusion. Moreover, performance is close to the oracle, thereby suggesting TE-CDE can be used to obtain a useful measure of uncertainty. This can then be used to exclude uncertain samples and improve performance. An important observation from a clinical perspective is that a small percentage ($< 10\%$) of samples are responsible for the majority of the overall estimation error. This highlights the utility that by deferring uncertain samples to clinicians we could significantly improve overall predictive and patient outcomes.
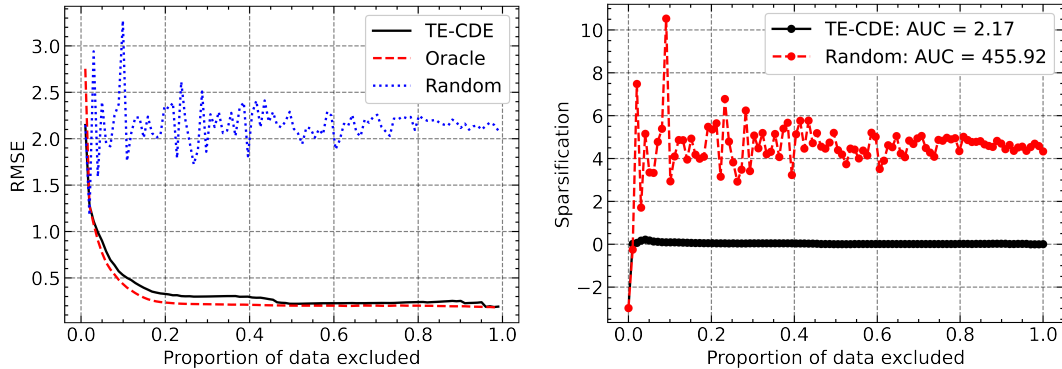
(a) Exclusion based on uncertainty: decreased error as more uncertain samples are excluded. The result highlights that a small proportion of samples are responsible for most of the error

(b) Sparsification curve with AUSE

Figure 8: Evaluation of the uncertainty representations for TE-CDE based on exclusions and sparsification

F.2. Data Efficiency Benchmark for Different Methods

Data efficiency is critical in clinical scenarios, where we may not have large datasets. Hence, we wish to quantify the sensitivity of the different methods to the size of the training set.

We contrast the RMSE for counterfactual estimation of TE-CDE, CRN and RMSN for varying amounts of training data, $n = 1000, 5000$ relative to $n = 10000$. We fix the level of time-varying confounding as $\gamma = 2$ and $\kappa = 10$.

The results in Table 5 highlight that TE-CDE achieves the lowest RMSE percentage drop-off for all amounts of training data. Whilst, all methods demonstrate reduced performance as the amount of training data decreases, the performance degradation is less severe for TE-CDE compared to CRN and more so when compared to RMSN. This demonstrates that TE-CDE is more data efficient, which is valuable in clinical settings where the data volume may be limited.

We posit the reason for this phenomenon is that since TE-CDE operates in continuous-time and has a z_t defined over all time points, that less data is required to learn the underlying dynamics. Whereas, both CRN and RMSN in the discrete paradigm would need more "discrete" examples in order to learn the underlying dynamics for different time steps.

Table 5: Data efficiency of models trained on varying amounts of training samples for fixed $\gamma = 2$

Train samples	Model	% Performance reduction (RMSE) vs. 10000 train samples (lower better)
	TE-CDE	**4.8%**
5000	CRN	6.4%
	RMSN	25.6%
	TE-CDE	**17.1%**
1000	CRN	18.9%
	RMSN	57.2%

F.3. Latent Representation and Discovery

F.3.1. TREATMENT INVARIANT LATENT REPRESENTATION

We wish to evaluate whether the latent representation z_t over time has indeed learnt treatment invariant representations. Figure 9 illustrates via low dimensional T-SNE embeddings that the TE-CDE latent state learns treatment invariant representations for different treatments over different time periods i.e. with no treatment, chemotherapy, radiotherapy, combined chemotherapy and radiotherapy.
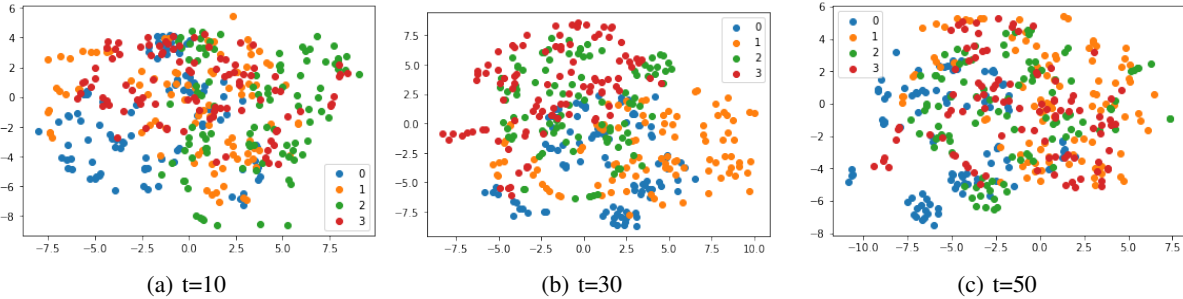

(a) t=10   (b) t=30   (c) t=50

Figure 9: T-SNE low-dimension embedding of the latent representation of TE-CDE, i.e z_t at different timesteps. It highlights that the representation is invariant of treatment, hence assists to deal with time-dependent confounding originating from treatment history H_t. The different color dots represent different treatment options

F.3.2. DISCOVERY AND INSIGHTS FROM LATENT REPRESENTATIONS

As demonstrated in Appendix F.1, we can ascertain which samples we are most uncertain about and use that to exclude/defer those samples. Additionally, it was highlighted that a small proportion of samples are responsible for the overall error.

From a clinical discovery perspective it is useful to ascertain unique characteristics of the uncertain samples and from a representation perspective whether there is a difference between the certain and uncertain samples. For the experiment we obtain the most certain and uncertain samples as characterized from the test set. Additionally, we obtain the most uncertain samples, which contributed most to the error from Figure 8(a). Understanding the nature of these samples is clinically beneficial.

We perform a T-SNE embedding (Van der Maaten & Hinton, 2008) of the latent representation (z) for the certain and uncertain samples at the final time-step. As illustrated in Figure 10, there is a marked difference between the certain and uncertain samples. Additionally, when comparing the mean tumor volumes for each group (i.e. indicative of severity), it is evident that the uncertain group consisted of more severe patients when compared to the certain group. Furthermore, the uncertain group had both greater mean and greater variability, where Uncertain volume: 5.37 ± 20 cm^3 and Certain volume: 2.57 ± 11 cm^3. These results highlight the distinct difference between the certain and uncertain samples. This demonstrates that not only do the latent representations capture this difference and that the uncertainty estimates are useful, but also that we can glean clinically useful information from a discovery perspective from the latent representation of TE-CDE.
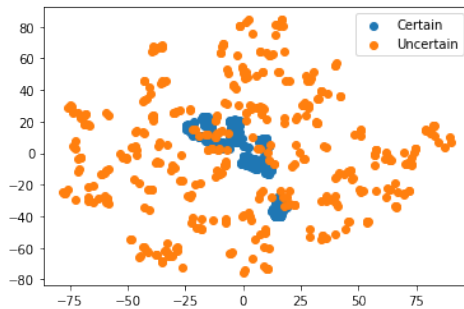

Figure 10: T-SNE of the latent representation for both certain (blue) and uncertain (orange) samples

F.4. Linear f and g

We conduct an ablation where f and g are linear instead of neural networks. We then assess performance for different amounts of time-dependent confounding ($\gamma = 2, 4, 6, 8, 10$). In terms of performance, we note a drop in RMSE of approximately 0.093 ± 0.056.

This of course highlights the importance of the neural network to performance. However, the model still retains performant properties suggesting the main source of performance gain is indeed from the CDE component.

F.5. Other assessment

We perform an added experiment based on the MIMIC-III ICU dataset (Johansson et al., 2016). Please note the caveat we list below. Similar to Bica et al. (2020b), we predict white blood cell count in response to antibiotic treatment. Consequently, we follow the related work and test on factuals. The results show similar performance for TE-CDE and CRN (difference ¡0.01 normalized RMSE), with both models predicting factuals with high accuracy. Note such experiments shouldn't be used to compare models for counterfactual prediction. As for the task on real data, we do not observe counterfactuals, rather only factuals. Hence, as discussed in Section 1, this is more a time-series evaluation rather than a counterfactual outcomes evaluation; which is the focus of this paper.

F.6. Impact of Time-Dependent Confounding Across Varying Sampling Intensities

Table 6 shows detailed results for the counterfactual predictions under varying sampling intensity (κ), across various levels of time-dependent confounding (γ). We compare TE-CDE to CRN, RMSN and also highlight results for the GP.

Table 6: Detailed results for counterfactual estimation under time-dependent confounding for different κ

Sampling intensity κ	Model	$\gamma = 2$	$\gamma = 4$	$\gamma = 6$	$\gamma = 8$	$\gamma = 10$
	TE-CDE	0.96 ± 0.01	1.55 ± 0.02	1.83 ± 0.08	2.26 ± 0.06	3.07 ± 0.35
	CRN	1.07 ± 0.09	1.63 ± 0.10	2.11 ± 0.29	3.23 ± 0.36	4.49 ± 0.38
$\kappa = 1$	RMSN	1.92 ± 0.07	2.42 ± 0.11	2.66 ± 0.15	4.00 ± 0.12	4.16 ± 0.11
	TE-CDE ($\mu = 0$)	1.04 ± 0.16	1.77 ± 0.11	2.14 ± 0.34	3.00 ± 0.21	4.26 ± 0.51
	GP	1.74 ± 0.76	2.80 ± 0.93	3.74 ± 0.40	6.15 ± 1.14	6.97 ± 1.02
	TE-CDE	0.88 ± 0.01	1.31 ± 0.02	2.32 ± 0.08	2.62 ± 0.06	3.02 ± 0.35
	CRN	1.27 ± 0.15	1.78 ± 0.14	3.08 ± 0.21	4.11 ± 0.38	4.90 ± 0.31
$\kappa = 5$	RMSN	2.69 ± 0.12	2.92 ± 0.24	3.16 ± 0.18	4.265 ± 0.13	5.40 ± 0.16
	TE-CDE ($\mu = 0$)	0.88 ± 0.07	1.68 ± 0.05	2.97 ± 0.20	4.00 ± 0.22	4.64 ± 0.59
	GP	9.08 ± 1.53	14.19 ± 1.73	24.47 ± 2.89	39.89 ± 1.02	42.99 ± 2.76
	TE-CDE	0.78 ± 0.01	1.16 ± 0.02	1.93 ± 0.08	2.67 ± 0.06	3.03 ± 0.35
	CRN	1.31 ± 0.10	1.51 ± 0.11	3.10 ± 0.26	4.28 ± 0.29	4.84 ± 0.31
$\kappa = 10$	RMSN	1.60 ± 0.11	2.04 ± 0.12	3.67 ± 0.22	4.62 ± 0.27	5.41 ± 0.31
	TE-CDE ($\mu = 0$)	0.85 ± 0.06	1.23 ± 0.07	2.48 ± 0.06	3.49 ± 0.51	4.94 ± 0.59
	GP	9.41 ± 0.98	14.43 ± 0.95	24.74 ± 1.97	40.21 ± 0.93	43.78 ± 3.11

F.7. Treatment-Conditioned Sampling

Table 7 shows detailed results for counterfactual estimation for treatment conditioned sampling (including results for the GP). Sampling intensity κ is conditioned on treatment such that $\kappa_{treated} = 10$ and $\kappa_{untreated} = 1$ and $\gamma = 4$.

Table 7: Comparison of different methods TE-CDE, CRN, RMSN, and GP for counterfactual estimation with κ conditioned on treatment, where $\kappa_{treated} = 10$ and $\kappa_{untreated} = 1$.

Model	Overall	Treated	Untreated
TE-CDE	1.18 ± 0.05	1.56 ± 0.06	0.22 ± 0.02
CRN	1.57 ± 0.06	1.97 ± 0.05	0.64 ± 0.08
RMSN	3.06 ± 0.09	3.12 ± 0.07	2.83 ± 0.08
GP	11.05 ± 0.23	16.05 ± 0.54	5.66 ± 0.45

F.8. Forecasting at Additional Time Horizons

Figure 11 (a)-(c) below shows further results for forecasting at additional time horizons. The results demonstrate the same relative performance of different methods as was shown in the main paper.

(a) Horizon = t_{k+3}

(b) Horizon = t_{k+4}
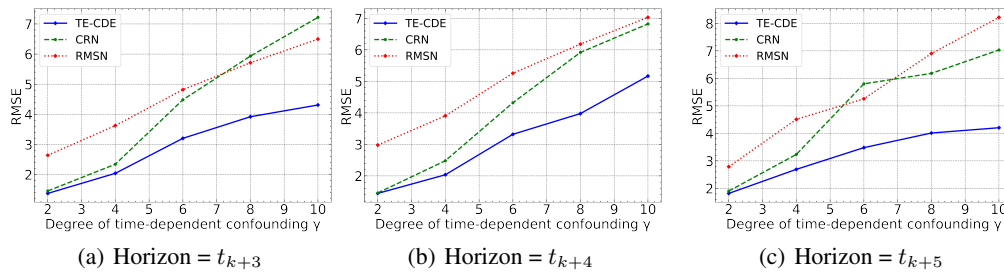
(c) Horizon = t_{k+5}

Figure 11: Results for counterfactual estimation at additional forecasting time horizons