Observable Neural ODEs for Identifiable Causal Forecasting in Continuous Time

Jennifer Wendland
wendland@uni-koblenz.de

Nicolas Freitag
nfreitag@uni-koblenz.de

Maik Kschischo
kschischo@uni-koblenz.de
Department of Computer Science
University of Koblenz
53424 Koblenz, Germany

Abstract

Causal inference in continuous-time sequential decision problems is challenged by hidden confounding and partially observed states. We show that, under explicit structural assumptions, observability of the latent state enables identification of dynamic treatment effects through a continuous-time conditional front-door adjustment, even in the presence of hidden confounding.

We derive a general adjustment formula and show that it reduces to a tractable state-space formula when unobserved contemporaneous disturbances are temporally uncorrelated. This formula expresses potential-outcome distributions under alternative treatment trajectories through the measurement model, latent dynamics, and the filtering distribution over latent states.

We propose Observable Neural ODEs (ObsNODEs), Neural ODE models in observable normal form that implement this tractable adjustment for causal forecasting. ObsNODEs learn continuous-time dynamics with states reconstructible from observations, enabling outcome prediction under alternative treatment paths.

Experiments on synthetic, semi-synthetic, and real-world clinical data demonstrate strong performance over recent sequence models, including external validation.

1 Introduction

Sequential decision-making requires predicting how actions affect future outcomes under evolving observations. In medicine, this arises in dynamic treatment regimes (DTRs), where treatments are repeatedly adapted to a patient's longitudinal history (Murphy, 2003; Robins, 2004).

A central challenge is the identification of causal treatment effects in continuous time. In observational data, time-dependent confounding and treatment–confounder feedback arise when covariates affect both treatment assignment and outcomes while themselves being affected by earlier treatments (Hernán, Robins, 2023). Identification becomes even more difficult when relevant components of the evolving disease state are latent and hidden variables affect both treatment and outcomes.

We address this problem from a state-space perspective. Under explicit structural assumptions, we show that observability makes the latent state reconstructible from treatment-outcome histories and thereby enables identification through a continuous-time conditional front-door adjustment, even in the presence of hidden confounding. We derive a general adjustment formula and show that, for temporally uncorrelated contemporaneous disturbances, it reduces to a tractable state-space formula.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Based on this result, we introduce *ObsNODEs*, Neural ODE models in observable normal form for causal forecasting under discrete and continuous treatment trajectories.

Contributions. (i) We link control-theoretic observability to causal identifiability in latent continuous-time state-space models and derive a general conditional front-door adjustment for dynamic treatment effects under hidden confounding. (ii) We derive a tractable state-space adjustment under temporally uncorrelated contemporaneous disturbances and propose *ObsNODEs*, which enforce observability by construction for causal forecasting under dynamic treatment regimes. (iii) We demonstrate strong causal forecasting performance on synthetic, semi-synthetic, and real-world clinical data compared with recent sequence models, including external validation.

2 Problem statement

We assume hat we observe a continuous-time multivariate outcome process $\boldsymbol{Y}_t$ and a time-varying treatment process $\boldsymbol{A}_t$ for times $t \in [0, t_c)$. Given the observed history $\{\boldsymbol{a}_{[0,t_c)}, \boldsymbol{y}_{[0,t_c]}\}$ with $\boldsymbol{a}_{[0,t_c)} = \{\boldsymbol{A}_t = \boldsymbol{a}_t | 0 \leq t < t_c\}$ and $\boldsymbol{y}_{[0,t_c]} = \{\boldsymbol{Y}_t = \boldsymbol{y}_t | 0 \leq t \leq t_c\}$, the goal is to predict the probability density of the conditional potential outcome $\boldsymbol{Y}_t(\boldsymbol{a}_{[t_c,t_c+s)})$ on $[t_c, t_c + s)$ under a hypothetical treatment path $\boldsymbol{a}_{[t_c,t_c+s)} = \{\boldsymbol{A}_t = \boldsymbol{a}_t | t_c \leq t < t_c + s\}$. More precisely, we seek to identify the conditional potential-outcome density (CPOD) $P_{\boldsymbol{Y}_{t_c+s}(\boldsymbol{a}_{[t_c,t_c+s)})|\boldsymbol{Y}_{t_c},\boldsymbol{O}_{[0,t_c)}}(\boldsymbol{y}_{t_c+s}|\boldsymbol{y}_{t_c}, \boldsymbol{o}_{[0,t_c)})$, given the last observation $\boldsymbol{y}_{t_c}$ and the history of observations and treatments $\boldsymbol{o}_{[0,t_c)} = \{\boldsymbol{a}_{[0,t_c)}, \boldsymbol{y}_{[0,t_c)}\}$.

In clinical applications, $\boldsymbol{Y}_t$ may contain irregularly sampled laboratory values, vital signs, or imaging embeddings, while $\boldsymbol{A}_t$ represents binary, categorical, or continuous treatments. Outside randomized trials, treatment assignment depends on the evolving patient history, which induces time-dependent confounding and treatment-confounder feedback.

We treat all measured variables as components of $\boldsymbol{Y}_t$ and define outcomes, when needed, as functions of $\boldsymbol{Y}_t$. This entails no loss of generality for the identification results below. We use outcome and observed process interchangeably for $\boldsymbol{Y}_t$.

3 Preliminaries

3.1 Nonlinear state-space model

A continuous-time deterministic state-space model is given by

$$\dot{\boldsymbol{z}}_t = \boldsymbol{f}(\boldsymbol{z}_t, \boldsymbol{a}_t) \quad \text{(1a)}, \quad \boldsymbol{y}_t = \boldsymbol{h}(\boldsymbol{z}_t) \quad \text{(1b)}, \quad \boldsymbol{z}_0 = \boldsymbol{\zeta} \quad \text{(1c)} \tag{1}$$

where $\boldsymbol{z}_t \in \mathbb{R}^{d_z}$ is a latent state, $\boldsymbol{a}_t \in \mathbb{R}^{d_a}$ a treatment or control, and $\boldsymbol{y}_t \in \mathbb{R}^{d_y}$ the observed process. The measurement function $\boldsymbol{h}$ is not necessarily invertible, in particular when $d_z > d_y$. We assume $\boldsymbol{f}$ is Lipschitz in $\boldsymbol{z}$, so the system admits a unique solution for any admissible control path and initial condition.

For a given control path, observability asks whether the latent state or, equivalently, its initial value, is uniquely determined by the observed output over a finite horizon (Bernard, 2019). For a given control $\boldsymbol{a}$ and initial condition $\boldsymbol{z}_0 = \boldsymbol{\zeta}$, let $\boldsymbol{z}_t(\boldsymbol{a}, \boldsymbol{\zeta})$ denote the resulting state trajectory, and define the corresponding output trajectory by $\boldsymbol{y}_t(\boldsymbol{a}, \boldsymbol{\zeta}) := \boldsymbol{h}(\boldsymbol{z}_t(\boldsymbol{a}, \boldsymbol{\zeta}))$.

Definition 3.1. The system (1) is **uniformly observable** on an open set $U \subset \mathbb{R}^{d_z}$ if, for any control path $\boldsymbol{a} : [0, T] \to \mathbb{R}^{d_a}$ and any $\boldsymbol{\zeta}, \boldsymbol{\eta} \in U$,

$$\boldsymbol{y}_t(\boldsymbol{a}, \boldsymbol{\zeta}) = \boldsymbol{y}_t(\boldsymbol{a}, \boldsymbol{\eta}) \ \ \forall t \in [0, T] \quad \Longrightarrow \quad \boldsymbol{\zeta} = \boldsymbol{\eta},$$

where $[0, T]$ is the maximal interval on which both solutions exist.

3.2 Potential outcomes and graphical causal models

We use the potential-outcomes framework and graphical causal models to formalize interventions (Robins, 1986; Pearl, 2000; Peters et al., 2017; Hernán, Robins, 2023). We assume consistency, the stable unit treatment value assumption (SUTVA), and positivity. SUTVA excludes interference between units and multiple versions of treatment (Rubin, 1980; Hernán, Robins, 2023), while positivity

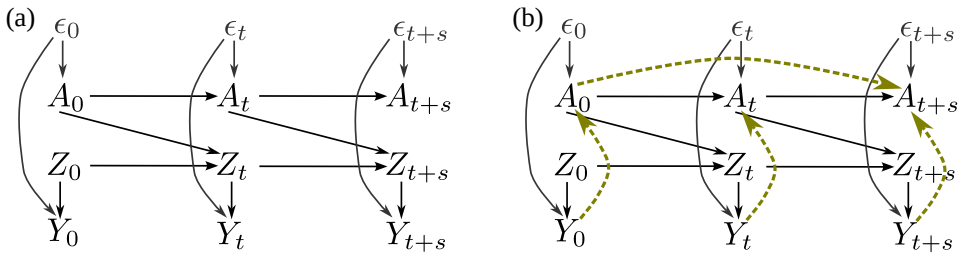


Figure 1: (a) Temporal DAG for treatment A, observation Y, and latent state Z at three consecutive time points. The temporally independent latent disturbances ϵ jointly affect contemporaneous treatment assignment and observations. (b) Under a dynamic treatment regime, treatment additionally depends on the observed treatment-outcome history, as indicated by the green dashed arrows.

requires that the considered interventions lie within the support of the observational treatment process conditional on the relevant history. In longitudinal observational settings, identifiability of causal treatment effects is complicated by time-dependent confounding and treatment-confounder feedback.

3.2.1 The conditional front-door criterion

To connect causal identification with latent state-space models, we use the conditional front-door criterion (Xu et al., 2024), an extension of the front-door criterion (Pearl, 2000) for estimating the effect of an intervention in the presence of an unobserved common cause of both treatment and outcome:

Theorem 3.2. *(Xu et al., 2024) A set of variables M fulfills the **conditional front-door criterion (CFD)** relative to the ordered pair of variables (A, Y) in a directed acyclic graph (DAG) $\mathcal{G}$, if (i) M intercepts all directed paths from A to Y, (ii) there exists a set of conditioning variables W that blocks all backdoor paths from A to M and (iii) all backdoor paths between M and Y are blocked by $\{A\} \cup W$. If M fulfills the CFD relative to (A, Y) with conditioning variables W the causal effect estimate is given by*

$$P(Y(a)) = \sum_{m,w,a'} P(y|a', m, w)P(a'|w)P(m|a, w)P(w), \tag{2}$$

where the sum over a' runs over alternative treatments.

4 Observable Neural ODEs for dynamic treatment regimes

We model the observed process by a latent Neural ODE (Chen et al., 2018; Rubanova et al., 2019; Kidger et al., 2020) with parameters θ

$$\boldsymbol{Z}_0 \sim p_\theta(\boldsymbol{Z}_0), \qquad \dot{\boldsymbol{Z}}_t = \boldsymbol{f}_\theta(\boldsymbol{Z}_t, \boldsymbol{A}_t), \qquad \boldsymbol{Y}_t \sim q_\theta(\cdot|\boldsymbol{Z}_t), \tag{3}$$

where p_θ parameterizes the initial state distribution and $\boldsymbol{f}_\theta$ the vector field. The state is mapped via the dsitribution $q(\cdot|\boldsymbol{Z}_t, \boldsymbol{\epsilon}_t)$ to the observations and the observation model q_θ describes the marginal observation distribution. We interpret the treatment process as an intervention on the latent dynamics, as illustrated in Fig. 1(a). Typically, the treatment is adjusted according to the observations and we have a dynamic treatment regime (Murphy, 2003; Robins, 2004; Hernán, Robins, 2023): $\boldsymbol{A}_t = \boldsymbol{\pi}(\boldsymbol{y}_t, \boldsymbol{y}_{[0,t)}, \boldsymbol{a}_{[0,t)})$ (see Fig. 1(b)). The temporally independent latent disturbance $\boldsymbol{\epsilon}_t$ may additionally affect both contemporaneous treatment assignment and observation, as illustrated in Fig. 1.

4.1 Causal identifiability for the state-space model

To establish an adjustment formula for the CPOD, we make the following assumptions for the data generating process, see Fig. 1.

Assumption 4.1. The latent state $\boldsymbol{Z}_t$ is a controlled continuous-time Markov process with transition probability density $p(\boldsymbol{z}_{t+s}|\boldsymbol{a}_{[t,t+s)}, \boldsymbol{z}_t)$.

Assumption 4.2. Given the latent state, the observation is independent of previous observations and treatments: $p(\boldsymbol{y}_t|\boldsymbol{z}_t, \boldsymbol{y}_{[0,t)}, \boldsymbol{a}_{[0,t)}) = p(\boldsymbol{y}_t|\boldsymbol{z}_t)$.

Theorem 4.3. *Under consistency, SUTVA, overlap (positivity), Assumptions 4.1 and 4.2, and observability of the latent state, the conditional potential-outcome distribution under $\boldsymbol{a}_{[t,t+s)}$ is identified by the adjustment formula*

$$P_{\boldsymbol{Y}_{t+s}(\boldsymbol{a}_{[t,t+s)})|\boldsymbol{Y}_t,\boldsymbol{O}_{[0,t)}}(\boldsymbol{y}_{t+s}|\boldsymbol{y}_t,\boldsymbol{o}_{[0,t)}) = \int_{\boldsymbol{z}_{t+s}}\int_{\boldsymbol{z}_t} p(\boldsymbol{z}_t|\boldsymbol{y}_t,\boldsymbol{o}_{[0,t)})\, p(\boldsymbol{z}_{t+s}|\boldsymbol{a}_{[t,t+s)},\boldsymbol{z}_t)\, p(\boldsymbol{y}_{t+s}|\boldsymbol{z}_{t+s}), \tag{4}$$

This identification remains valid when temporally independent latent disturbances $\boldsymbol{\epsilon}_t$ induce unobserved contemporaneous dependence between treatment assignment and observations. The result also applies to dynamic treatment regimes under sequential positivity, with the transition density induced by the regime (see Fig. 1(b)).

Remark 4.4. Assumption 4.2 holds when $\boldsymbol{\epsilon}_t$ is an exogenous, temporally independent disturbance. With serially correlated hidden disturbances, the causal effect remains identifiable through the general conditional front-door adjustment, but the formula no longer reduces to (4); see (13) in Appendix A.1.

Proof. We derive a discrete-time version of the adjustment formula (4) in Theorem 4.3 and relegate details for the continuous-time version to Appendix A.1. Consider three consecutive time points $0, t$ and $t+s$. A time-discretized approximation of the ODE in (3) is given by $\boldsymbol{Z}_{t+s} \approx \boldsymbol{Z}_t + \boldsymbol{f}(\boldsymbol{Z}_t, \boldsymbol{A}_t)\, s$. Assume we have observed $\boldsymbol{o}_0 = \{\boldsymbol{a}_0, \boldsymbol{y}_0\}$ and $\boldsymbol{y}_t$. To apply the CFD in Theorem 3.2 we make the following identifications

$$A \leftarrow\!\shortmid \boldsymbol{A}_t, \quad Y \leftarrow\!\shortmid \boldsymbol{Y}_{t+s}, \quad M \leftarrow\!\shortmid \boldsymbol{Z}_{t+s}, \quad W \leftarrow\!\shortmid \{\boldsymbol{Z}_t, \boldsymbol{Y}_t, \boldsymbol{A}_0, \boldsymbol{Y}_0\} = \{\boldsymbol{Z}_t, \boldsymbol{Y}_t, \boldsymbol{O}_0\}\,.$$

Although $\boldsymbol{Z}_t$ is latent, observability in this discrete-time representation implies that it can be uniquely reconstructed from the treatment-outcome history. By Assumption 4.1, the mediator distribution under treatment $\boldsymbol{a}_t$ is then given by $p(\boldsymbol{z}_{t+s}|\boldsymbol{a}_t, \boldsymbol{z}_t)$. Consequently, both the conditioning state and the mediator distribution required by the CFD adjustment are identifiable.

We can check graphically in Fig. 1(a,b) that $M = \boldsymbol{Z}_{t+s}$ satisfies the CFD criterion relative to the ordered pair $(A, Y) = (\boldsymbol{A}_t, \boldsymbol{Y}_{t+s})$, if we use $W = \{\boldsymbol{Z}_t, \boldsymbol{Y}_t, \boldsymbol{O}_0\}$ as conditioning set. Now we compute the probability density for the potential outcome $\boldsymbol{Y}_{t+s}(\boldsymbol{a}_t)$ when applying treatment $\boldsymbol{A}_t = \boldsymbol{a}_t$ at time t using (2) and replacing the sums by integrals

$$P_{\boldsymbol{Y}_{t+s}(\boldsymbol{a}_t)}(\boldsymbol{y}_{t+s}) = \int_{\boldsymbol{z}_{t+s}}\int_{\boldsymbol{z}_t,\boldsymbol{y}_t,\boldsymbol{o}_0}\int_{\boldsymbol{a}'_t} \{p(\boldsymbol{y}_{t+s}|\boldsymbol{a}'_t, \boldsymbol{z}_{t+s}, \boldsymbol{z}_t, \boldsymbol{y}_t, \boldsymbol{o}_0) \tag{5}$$

$$\cdot\, p(\boldsymbol{a}'_t|\boldsymbol{z}_t, \boldsymbol{y}_t, \boldsymbol{o}_0)\; p(\boldsymbol{z}_{t+s}|\boldsymbol{a}_t, \boldsymbol{z}_t, \boldsymbol{y}_t, \boldsymbol{o}_0)\; p(\boldsymbol{z}_t, \boldsymbol{y}_t, \boldsymbol{o}_0)\}\,. \tag{6}$$

By Assumptions 4.2 and 4.1 we have the following independence relations

$$p(\boldsymbol{y}_{t+s}|\boldsymbol{a}'_t, \boldsymbol{z}_{t+s}, \boldsymbol{z}_t, \boldsymbol{y}_t, \boldsymbol{o}_0) = p(\boldsymbol{y}_{t+s}|\boldsymbol{z}_{t+s})$$
$$p(\boldsymbol{z}_{t+s}|\boldsymbol{a}_t, \boldsymbol{z}_t, \boldsymbol{y}_t, \boldsymbol{o}_0) = p(\boldsymbol{z}_{t+s}|\boldsymbol{a}_t, \boldsymbol{z}_t)\,.$$

Integrating over $\boldsymbol{a}'_t$ and using $p(\boldsymbol{z}_t, \boldsymbol{y}_t, \boldsymbol{o}_0) = p(\boldsymbol{z}_t|\boldsymbol{y}_t, \boldsymbol{o}_0)\; p(\boldsymbol{y}_t, \boldsymbol{o}_0)$ we find

$$P_{\boldsymbol{Y}_{t+s}(\boldsymbol{a}_t)}(\boldsymbol{y}_{t+s}) = \int_{\boldsymbol{z}_{t+s}}\int_{\boldsymbol{z}_t,\boldsymbol{y}_t,\boldsymbol{o}_0} \{p(\boldsymbol{y}_{t+s}|\boldsymbol{z}_{t+s})\, p(\boldsymbol{z}_{t+s}|\boldsymbol{a}_t, \boldsymbol{z}_t)\; p(\boldsymbol{z}_t|\boldsymbol{y}_t, \boldsymbol{o}_0)\, p(\boldsymbol{y}_t, \boldsymbol{o}_0)\}\,,$$

from which we read the conditional density for the potential outcome $\boldsymbol{Y}_{t+s}(\boldsymbol{a}_t)$ when applying treatment $\boldsymbol{A}_t = \boldsymbol{a}_t$ at time t, given we have observed outcome $\boldsymbol{y}_t$ at time t and the treatment-outcome history $\boldsymbol{o}_0$ at time zero

$$P_{\boldsymbol{Y}_{t+s}(\boldsymbol{a}_t)|\boldsymbol{Y}_t,\boldsymbol{O}_0}(\boldsymbol{y}_{t+s}\,|\boldsymbol{y}_t, \boldsymbol{o}_0) = \int_{\boldsymbol{z}_{t+s}}\int_{\boldsymbol{z}_t} \{p(\boldsymbol{y}_{t+s}|\boldsymbol{z}_{t+s})\, p(\boldsymbol{z}_{t+s}|\boldsymbol{a}_t, \boldsymbol{z}_t)\; p(\boldsymbol{z}_t|\boldsymbol{y}_t, \boldsymbol{o}_0)\}\,.$$

The adjustment formula remains valid for dynamic treatment regimes in which the current observation $\boldsymbol{y}_t$ and the observed history $\boldsymbol{o}_0$ determine the treatment $\boldsymbol{a}_t$, because the same CFD criterion continues to hold (Fig. 1(b)). □

The key requirement is that the latent state is observable from the observed history. For the deterministic ODE in (3) the dynamics is given by the point mass: $p(\boldsymbol{z}_{t+s}|\boldsymbol{a}_{[t,t+s)}, \boldsymbol{z}_t) = \delta(\boldsymbol{z}_{t+s} - \boldsymbol{z}_t - \int_t^{t+s} \boldsymbol{f}(\boldsymbol{z}_\tau, \boldsymbol{a}_\tau)\, d\tau)$. If we replace the ODE by a stochastic differential equation we need a notion of stochastic observability (Van Handel, 2009).

4.2 Neural ODEs in observable normal form (ObsNODEs)

The identification formula (4) requires that the latent state can be inferred from observed outcomes and treatment history. Standard latent Neural ODEs do not guarantee this property. We therefore parameterize the latent dynamics in continuous triangular observable normal form (Bernard, 2019; Adamy, 2024):

$$\dot{\boldsymbol{Z}}_t = \begin{pmatrix} \dot{\boldsymbol{Z}}_t^{(1)} \\ \vdots \\ \dot{\boldsymbol{Z}}_t^{(i)} \\ \vdots \\ \dot{\boldsymbol{Z}}_t^{(m)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{Z}_t^{(2)} + \boldsymbol{\phi}_{\theta_1}^{(1)}(\boldsymbol{Z}_t^{(1)}, \boldsymbol{a}_t) \\ \vdots \\ \boldsymbol{Z}_t^{(i+1)} + \boldsymbol{\phi}_{\theta_i}^{(i)}(\boldsymbol{Z}_t^{(1)}, \ldots, \boldsymbol{Z}_t^{(i)}, \boldsymbol{a}_t) \\ \vdots \\ \boldsymbol{\phi}_{\theta_m}^{(m)}(\boldsymbol{Z}_t, \boldsymbol{a}_t) \end{pmatrix}, \qquad \boldsymbol{Y}_t = \boldsymbol{Z}_t^{(1)}. \tag{7}$$

Here the state $\boldsymbol{Z}_t = \left(\boldsymbol{Z}_t^{(1)}, \ldots, \boldsymbol{Z}_t^{(m)}\right) \in \mathbb{R}^{d_z}$ is partitioned into m blocks $\boldsymbol{Z}_t^{(i)} \in \mathbb{R}^{d_y}$, so that $d_z = m d_y$, and each $\boldsymbol{\phi}_{\theta_i}^{(i)}$ is implemented by a neural network. Under suitable conditions, observable systems can be transformed into this normal form without loss of expressiveness (Bernard, 2019).

We allow for additive observation noise, $\boldsymbol{Y}_t = \boldsymbol{Z}_t^{(1)} + \boldsymbol{\eta}_t$, where $\boldsymbol{\eta}_t$ is a centered noise process. This noise does not affect identification. The hidden disturbance ($\boldsymbol{\epsilon}_t$) need not be centered and is accounted for by the corresponding adjustment formula.

4.3 Prediction and training of ObsNODEs

We approximate the filtering distribution $p(\boldsymbol{z}_t | \boldsymbol{y}_t, \boldsymbol{o}_{[0,t)})$ with an RNN encoder, following (Chen et al., 2018). Given the inferred state $\boldsymbol{z}_t$, forecasts under a hypothetical treatment path $\boldsymbol{a}_{[t,t+s)}$ are obtained by integrating (7) forward in time and projecting to the observed space.

Training is self-supervised. For each unit, we choose a decision time t_c and provide observations up to t_c. The model then predicts future observations on $(t_c, t_f]$.

Observations are available only at irregular time points $\mathcal{T}_{ij}$. Let $\mathcal{T}_{ij}(t_c, t_f) = \mathcal{T}_{ij} \cap (t_c, t_f]$ denote the observation times of component j for unit i in that forecast window $(t_c, t_f]$. The predicted outcome vector for unit i and outcome component j at time t is to be compared against the corresponding observed value y_{tij}^{dat} in the training data. We minimize the masked squared error

$$\mathcal{L}_{(t_c,t_f]} = \sum_{i=1}^{n} \sum_{j=1}^{d_y} \frac{1}{n\,\sigma_j^2\,|\mathcal{T}_{ij}(t_c, t_f)|} \sum_{t \in \mathcal{T}_{ij}(t_c,t_f)} \left(y_{tij} - y_{tij}^{\mathrm{dat}}\right)^2,$$

where $|\mathcal{T}_{ij}(t_c, t_f)|$ denotes the number of observations of component j for unit i in the forecasting interval $(t_c, t_f]$. This normalization compensates for varying observation frequencies across units and outcome components. The scaling factor σ_j^2 is the empirical variance of the $j-$th outcome component across all units and observed time points in the training data and can be set to 1 for standardized data.

During training, a decision time $t_c \in [0, t_f)$ is selected for all units in a batch of size n and the corresponding losses $\mathcal{L}_{(t_c,t_f]}$ are averaged over different batches. We found that both a fixed set of possible values for t_c evenly distributed over $[0, t_f)$ or a uniform random choice led to very similar results. Choosing different decision times t_c in each batch exposes the model to varying assimilation times and forecast horizons.

5 Experiments

We evaluate ObsNODE on synthetic and semi-synthetic datasets where the ground-truth causal effect of a prescribed treatment path on the outcome trajectories is known and, in addition, on a real world data set with an external validation. We compare against the IGC-Net (Hess et al., 2026), SCIP-Net (Hess, Feuerriegel, 2025) and two Treatment Effect Neural Controlled Differential Equation models (Seedat et al., 2022): optAB for categorical treatments (Wendland et al., 2024) and doseAI for categorical and continuous treatments (Wendland, Kschischo, 2025).

Forecast root mean squared error (RMSE) was scaled with the standard deviation of each dynamic variable in the test set and is reported as a function of assimilation time (the duration of the treatment-observation history t_c) and prediction horizon (Figures 2–4). The RMSE values of these heatmaps are given in `https://github.com/JenniferJaschob/ObsNODE.git`.

As an ablation study, we compare ObsNODE with an otherwise identical latent Neural ODE without the observable normal form. For each test patient, both models assimilate the factual history up to $t_c = 5$ and predict outcomes under five prespecified treatment regimes. Counterfactual bias (CB) is the signed difference between predicted and ground-truth potential outcomes, averaged over all test patients and six prediction time points. We also evaluate the average treatment effect (ATE) of each regime relative to a baseline treatment (Tables 2 and 4).

Results are averaged over five independent runs with disjoint training, validation, and test sets. For synthetic and semi-synthetic data we vary the confounding strength γ. Details on data generation and the counterfactual treatment scenarios are provided in the Appendix A.4.1 and A.4.2.

5.1 Experiments with fully-synthetic data

Data: The synthetic dataset is generated from a dynamic model of tumor volume (Geng et al., 2017) and body weight (Wendland, Kschischo, 2025) under chemotherapy and radiotherapy. The treatment vector $\boldsymbol{a}_t$ represents chemo- and radiotherapy doses administered in monthly cycles, and the outcome $\boldsymbol{y}_t$ consists of tumor volume and body weight.

To simulate time-dependent confounding, treatment probabilities depend on the mean tumor volume over the previous 15 days (Melnychuk et al., 2022). Patient heterogeneity is introduced by sampling tumor growth and treatment response parameters from predefined distributions (Appendix A.4.1).

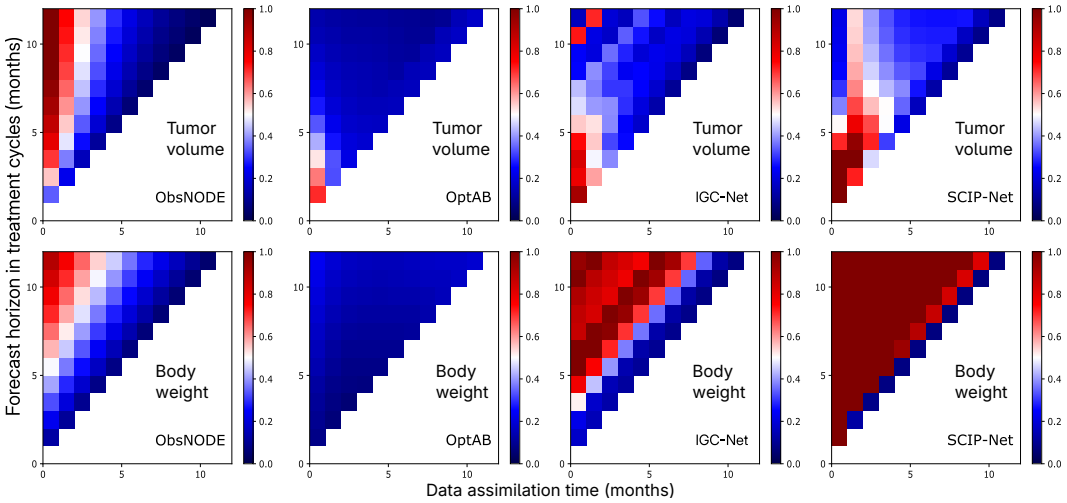


Figure 2: **Synthetic cancer dataset.** RMSE heatmaps (mean over five runs) for standardized (Z-score) tumor volume and body weight forecasts at confounding strength $\gamma = 4$. The horizontal axis shows the assimilation time and the vertical axis the forecast horizon. Columns correspond to ObsNODE, doseAI (Wendland, Kschischo, 2025), IGC-Net (Hess et al., 2026) and SCIP-Net (Hess, Feuerriegel, 2025). RMSE values larger than one were clipped at one.

Results: Figure 2 compares tumor volume and body weight forecast RMSE for $\gamma = 4$. For long horizons, doseAI achieves the lowest error, followed by ObsNODE and IGC-Net, whereas ObsNODE performs best for short horizons ($\tau \leq 2$ cycles). Detailed RMSE values for the tumor volume at different confounding strengths γ and prediction horizons τ are reported in Table 1. ObsNODE consistently shows the smallest variance across different runs.

ObsNODE consistently outperforms the Neural ODE (NODE) in CB and ATE estimation (Table 2), reducing the mean absolute ATE error from 0.5289 to 0.1723 for tumor volume and from 2.0289 to 0.0933 for body weight.

γ	MODEL/ s	1	2	3	4	5	6
4	OBSNODE	**0.06±0.00**	**0.11±0.00**	0.14±0.00	0.17±0.00	0.19±0.00	0.21±0.01
	DOSEAI	0.16±0.02	0.15±0.02	**0.12±0.01**	**0.12±0.01**	**0.11±0.00**	**0.10±0.01**
	IGC-NET	0.13±0.01	0.18±0.04	0.22±0.05	0.22±0.03	0.17±0.01	0.26±0.14
	SCIP-NET	0.16±0.00	0.27±0.02	0.32±0.04	0.30±0.05	0.28±0.06	0.27±0.07
	NODE	0.61±0.00	0.47±0.00	0.37±0.00	0.30±0.00	0.25±0.00	0.22±0.00
6	OBSNODE	**0.06±0.00**	**0.10±0.00**	0.13±0.00	0.15±0.00	0.17±0.00	0.18±0.01
	DOSEAI	0.12±0.02	0.11±0.01	**0.10±0.01**	**0.10±0.01**	**0.10±0.00**	**0.10±0.01**
	IGC-NET	0.12±0.01	0.18±0.03	0.20±0.04	0.20±0.03	0.16±0.01	0.25±0.15
	SCIP-NET	0.14±0.00	0.24±0.02	0.31±0.05	0.29±0.05	0.27±0.06	0.26±0.06
	NODE	0.56±0.00	0.43±0.00	0.34±0.00	0.28±0.00	0.24±0.00	0.20±0.00
8	OBSNODE	**0.05±0.00**	**0.09±0.00**	0.12±0.00	0.14±0.00	0.16±0.00	0.18±0.01
	DOSEAI	0.12±0.01	0.11±0.00	**0.11±0.00**	**0.11±0.01**	**0.12±0.01**	**0.13±0.02**
	IGC-NET	0.11±0.01	0.18±0.03	0.20±0.04	0.19±0.03	0.16±0.01	0.25±0.17
	SCIP-NET	0.14±0.00	0.24±0.02	0.31±0.04	0.29±0.05	0.28±0.06	0.27±0.06
	NODE	0.55±0.00	0.43±0.00	0.35±0.00	0.29±0.00	0.25±0.00	0.22±0.00

Table 1: **The RMSE for the tumor volume from synthetic data.** The RMSE between predicted and observed tumor volume (mean ± std over five runs) for a fixed data assimilation time of $t_c = 5$ and different forecast horizons $s = t_f - t_c$ (in monthly treatment cycles) at different levels of confounding γ.

METRIC	VARIABLE	MODEL / CF_k	1	2	3	4	5
CB	TUMOR VOLUME	OBSNODE	**0.1738**	**0.1739**	**0.1724**	**0.1708**	**0.1708**
		NODE	-0.5152	-0.5288	-0.5364	-0.5321	-0.5321
CB	BODY WEIGHT	OBSNODE	**0.0952**	**0.0941**	**0.0924**	**0.0924**	**0.0926**
		NODE	-2.0229	-2.0272	-2.0313	-2.0315	-2.0315
ATE	TUMOR VOLUME	TRUE	-0.0105	0.0031	0.0107	0.0063	0.0063
		OBSNODE	**0.1633**	**0.1770**	**0.1831**	**0.1771**	**0.1771**
		NODE	-0.5257	-0.5257	-0.5257	-0.5257	-0.5257
ATE	BODY WEIGHT	TRUE	-0.0016	0.0227	0.0068	0.0070	0.0070
		OBSNODE	**0.0937**	**0.0968**	**0.0992**	**0.0995**	**0.0995**
		NODE	-2.0245	-2.0245	-2.0245	-2.0245	-2.0245

Table 2: **Counterfactual bias (CB) and estimated average treatment effect (ATE)** relative to the baseline treatment for five treatment regimes (CF_1–CF_5) in the synthetic cancer dataset, with data assimilation time $t_c = 5$ and confounding level $\gamma = 4$. CB and ATE were averaged over six prediction time points, and all results are averaged over five runs.

5.2 Experiments with semi-synthetic data

Data: The second experiment employs a semi-synthetic dataset derived from real-world medical data. The MIMIC-IV dataset (Johnson et al., 2023) serves as the foundation, with a focus on the patient cohort diagnosed with sepsis. Based on this patient cohort, synthetic data are generated to represent patient trajectories and outcomes under both endogenous and exogenous dependencies, while explicitly incorporating treatment effects. The synthetic data generation methodology was originally defined using the MIMIC-III dataset (Schulam, Saria, 2018; Johnson et al., 2016; Melnychuk et al., 2022). Further details on the synthetic data generation process are provided in Appendix A.4.2.

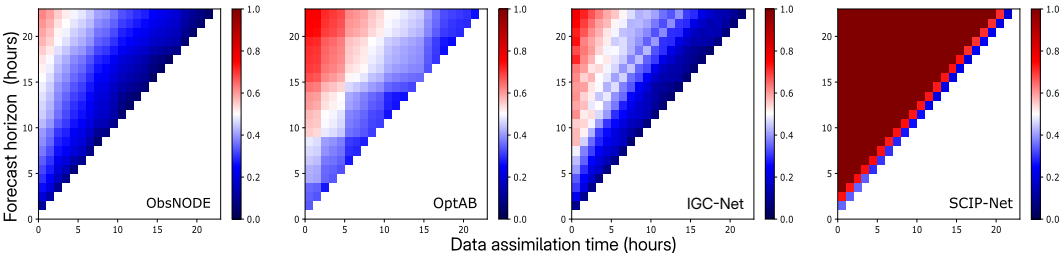


Figure 3: **Semi-synthetic MIMIC-IV sepsis dataset.** RMSE for ObsNODE, OptAB, IGC-Net and SCIP-Net as a function of data assimilation time (observation time) and forecast horizon over a 24-hour horizon. Results are averaged over five runs and RMSE values larger than one were clipped at one.

MODEL/ $s = t_f - t_c$	1	2	3	4	5	6
OBSNODE	0.13±0.00	**0.21±0.00**	**0.25±0.00**	**0.27±0.00**	**0.30±0.00**	**0.33±0.00**
OPTAB	0.34±0.04	0.35±0.04	0.41±0.02	0.42±0.02	0.44±0.02	0.45±0.02
IGC-NET	**0.10±0.03**	**0.21±0.01**	**0.25±0.01**	0.29±0.02	0.34±0.02	0.38±0.01
SCIP-NET	0.34±0.10	0.73±0.10	1.32±0.16	1.32±0.16	1.33±0.16	1.32±0.16
NODE	0.88±0.06	0.63±0.11	0.48±0.12	0.43±0.06	0.41±0.04	0.42±0.05

Table 3: **Semi-synthetic MIMIC-IV sepsis dataset.** RMSE (mean ± std over 5 runs) over a $s = t_f - t_c$ hour horizon when observing for a fixed data assimilation time of $t_c = 1$ hour.

METRIC	VARIABLE	MODEL / CF_k	1	2	3	4	5
CB	SEMI-SYNTHETIC OUTCOME	OBSNODE	**1.9165**	**1.2978**	**1.3962**	**2.6466**	**-0.0149**
		NODE	5.1987	2.4121	2.4121	7.6795	-0.5663
ATE	SEMI-SYNTHETIC OUTCOME	TRUE	-5.7596	-5.1849	-5.1849	-5.1849	-3.9458
		OBSNODE	**-3.8431**	**-3.8860**	**-3.7887**	**-3.9403**	**-3.7794**
		NODE	-0.5608	-2.7728	-2.7728	1.2556	-4.5121

Table 4: **Counterfactual bias (CB) and estimated average treatment effect (ATE)** relative to the baseline treatment for five treatment regimes (CF_1–CF_5) in the semi-synthetic MIMIC-IV sepsis dataset, with data assimilation time $t_c = 5$. CB and ATE were averaged over six prediction time points, and all results are averaged over five runs.

Results: Figure 3 shows results on the semi-synthetic MIMIC-IV sepsis dataset. ObsNODE achieves consistently a low RMSE across the 24-hour prediction horizon and performs better than OptAB, IGC-Net and SCIP-Net for longer horizons. RMSE statistics averaged over five runs (Table 3) show very small variance for ObsNODE, indicating stable predictions compared to the baselines.

For the semi-synthetic data, ObsNODE consistently achieves lower CB across all counterfactual treatment scenarios and reduces the mean absolute ATE error from 3.6537 to 1.4523 compared with the standard Neural ODE (NODE) without the observable normal form (Table 4).

5.3 Experiments with real-world data

Data: Selecting antibiotic therapy for sepsis patients is challenging (Wendland et al., 2024). We use the Sequential Organ Failure Assessment (SOFA) score, a summary measure of organ dysfunction, as the primary outcome and predict the effects of the antibiotics vancomycin, ceftriaxone, and piperacillin-tazobactam on the SOFA trajectory. In addition, we predict treatment effects on the side-effect markers creatinine, total bilirubin, and alanine aminotransferase (ALT). Further dynamic variables included in the observation vector are listed in Table 9 in Appendix A.4.3.

The MIMIC-IV sepsis cohort (Johnson et al., 2023) was used to train and test all models. In addition, ObsNODE was externally validated on the independent AmsterdamUMCdb sepsis cohort (Thoral et al., 2021).

Results: Table 5a summarizes the five-run robustness analysis. ObsNODE shows low variability across datasets sampled from the same distribution. Its competitive performance is particularly evident at longer forecast horizons, while IGC-Net performs better for very short-term predictions of selected laboratory variables (Figure 4).

Table 5b summarizes the external validation of ObsNODE, trained on MIMIC-IV, on AmsterdamUMCdb. Predictions were more accurate for creatinine and total bilirubin and remained comparable for SOFA score and ALT. The improved performance may reflect the higher observation density in AmsterdamUMCdb, which provides more measurements for data assimilation during prediction.

6 Related Work

Classical statistical approaches for longitudinal causal inference focus on estimating population-level average potential outcomes and do not condition on individual treatment and outcome histories (Robins, 1986; Bang, Robins, 2005; Lok, 2008; Van Der Laan, Gruber, 2012; Rytgaard et al., 2022). Consequently, they cannot be used for individualized outcome forecasting. These approaches

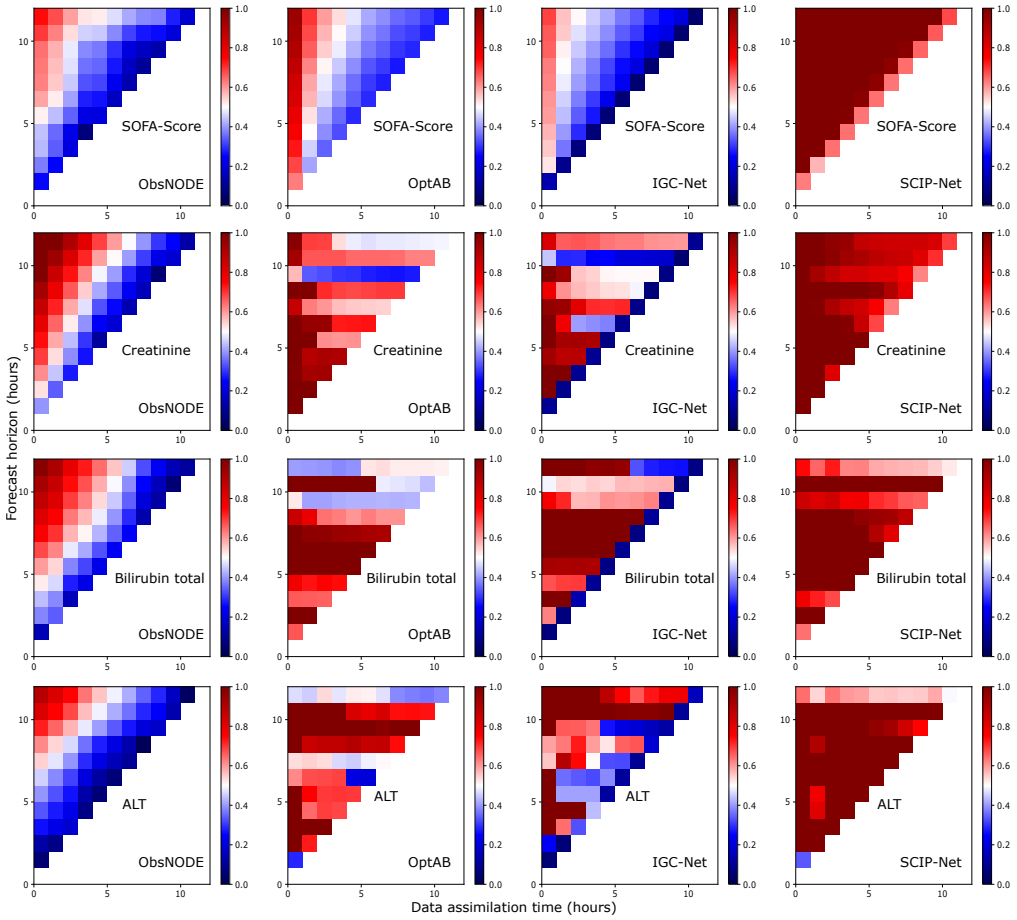


Figure 4: Forecasting performance on MIMIC-IV. Forecast RMSE for SOFA score, creatinine, total bilirubin, and alanine aminotransferase (ALT) across data assimilation times and forecast horizons.

are typically grounded in structural causal models and identification results such as the longitudinal g-formula. Extensions to continuous time include structural dynamic causal models (Bongers et al., 2022), which adopt a different formulation of interventions than considered here. Recent work has developed discrete-time sequence models for estimating conditional average potential outcomes (CAPO) (i.e. the expectation value of (4)), including recurrent and attention-based architectures (Bica et al., 2020; Li et al., 2020; Melnychuk et al., 2022; Hess et al., 2026). A limitation of these approaches is their reliance on fixed time discretizations, whereas clinical measurements are typically observed at irregular time points.

Continuous-time approaches based on neural differential equations address this limitation. Treatment Effect Controlled Differential Equations (TE-CDEs) (Seedat et al., 2022) combine controlled differential equations (Kidger et al., 2020) with representation balancing to mitigate time-dependent confounding. The original formulation considered binary treatments and was later extended to continuous treatments (Wendland, Kschischo, 2025). However, balancing-based approaches may not fully address time-dependent confounding in sequential settings (Melnychuk et al., 2022).

The SCIP-Net (Hess, Feuerriegel, 2025) is a recent continuous time method for estimating the CAPO based on a stabilized inverse propensity score weighting, which is, however, restricted to discrete treatment levels.

Our ObsNODE framework is most closely related to CF-ODE (Brouwer et al., 2022), which parameterizes a continuous-time state-space model of the form (3) using a Neural ODE. CF-ODE assumes that observations at treatment time uniquely determine the latent state, corresponding to the observability condition (Definition 3.1). However, this assumption is not structurally enforced and is difficult to verify in practice. In contrast, ObsNODE employs the observable normal form (7),

MODEL/ $s = t_f - t_c$	1	2	3	4	5	6
			SOFA-SCORE			
OBSNODE	0.24±0.00	**0.32±0.01**	**0.37±0.01**	**0.42±0.02**	0.47±0.03	**0.50±0.04**
OPTAB	0.41±0.01	0.46±0.01	0.50±0.02	0.53±0.02	0.57±0.02	0.59±0.02
IGC-NET	**0.09±0.00**	0.37±0.00	0.40±0.00	0.43±0.00	**0.45±0.00**	0.48±0.01
SCIP-NET	0.57±0.06	1.03±0.44	1.25±0.43	1.23±0.39	1.21±0.35	1.19±0.33
			CREATININE			
OBSNODE	0.34±0.00	**0.44±0.00**	**0.53±0.00**	**0.53±0.02**	**0.58±0.04**	0.65±0.07
OPTAB	1.14±0.02	0.93±0.02	0.89±0.03	1.23±0.03	0.96±0.01	**0.62±0.02**
IGC-NET	**0.09±0.02**	1.13±0.01	0.89±0.00	0.92±0.01	0.78±0.15	0.95±0.01
SCIP-NET	1.01±0.04	1.06±0.26	1.36±0.28	1.39±0.30	1.17±0.39	0.99±0.42
			BILIRUBIN-TOTAL			
OBSNODE	0.34±0.00	**0.38±0.00**	**0.46±0.02**	**0.54±0.04**	**0.62±0.07**	**0.70±0.11**
OPTAB	1.36±0.05	0.66±0.01	0.73±0.07	1.05±0.04	1.56±0.14	0.97±0.05
IGC-NET	**0.07±0.01**	1.40±0.02	0.69±0.01	0.91±0.00	1.07±0.01	1.59±0.03
SCIP-NET	1.23±0.17	0.72±0.12	1.13±0.11	1.24±0.12	1.84±0.14	1.17±0.14
			ALT			
OBSNODE	0.09±0.00	**0.22±0.01**	**0.27±0.03**	**0.33±0.06**	**0.39±0.10**	**0.47±0.13**
OPTAB	0.72±0.20	1.25±0.09	0.65±0.16	0.76±0.25	0.67±0.21	0.54±0.11
IGC-NET	**0.06±0.02**	0.64±0.17	1.34±0.21	0.41±0.02	0.36±0.03	0.90±0.23
SCIP-NET	1.75±0.20	1.51±0.30	0.81±0.28	0.77±0.37	2.24±0.45	1.45 ± 0.26

(a) RMSE for ObsNODE, OptAB, IGC-Net and SCIP-Net on a test set from the MIMIC-IV sepsis dataset.

VARIABLE/ $s = t_f - t_c$	1	2	3	4	5	6
SOFA-SCORE	0.26±0.00	0.37±0.01	0.43±0.02	0.44±0.02	0.51±0.02	0.59±0.03
CREATININE	**0.14±0.00**	**0.18±0.02**	**0.22±0.05**	**0.27±0.09**	**0.34±0.13**	**0.39±0.17**
BILIRUBIN-TOTAL	**0.01±0.00**	**0.11±0.02**	**0.18±0.05**	**0.25±0.09**	**0.33±0.13**	**0.40±0.17**
ALT	**0.01±0.00**	0.28±0.01	0.34±0.03	0.41±0.05	0.50±0.07	0.86±0.07

(b) RMSE for ObsNODE, trained on MIMIC-IV sepsis data and tested on an external validation set from AmsterdamUMCdb. Bold values indicate lower RMSE on AmsterdamUMCdb than on the MIMIC-IV test set reported above.

Table 5: **Forecasting results on the real-world Sepsis dataset.** RMSE for different prediction horizons (in hours) and a fixed data assimilation time of $t_c = 1$ hour. Mean and standard deviation are reported across five runs.

which guarantees observability by construction and makes explicit the link between control-theoretic observability and the identifiability condition required for the causal adjustment formula (4). Moreover, CF-ODE conditions only on prior observations and does not explicitly incorporate the treatment history.

Together, these limitations motivate a continuous-time state-space formulation that enforces observability and enables causal outcome prediction through an explicit adjustment formula, which we develop in the ObsNODE framework.

7 Conclusion and Limitations

We introduced Observable Neural ODEs (ObsNODEs), linking control-theoretic observability to causal identifiability in continuous-time state-space models. ObsNODEs enforce observability by construction, enable causal forecasting under alternative treatment trajectories, and perform strongly on synthetic, semi-synthetic, and clinical data, including external validation.

The approach relies on consistency, positivity, and the stated state-space assumptions. The tractable adjustment formula permits contemporaneous hidden confounding but assumes temporally independent disturbances; serially correlated confounding requires the general path-space adjustment derived in the Appendix. State reconstruction from finite, noisy observations remains approximate, and causal accuracy cannot be assessed on clinical data without counterfactual ground truth. Future work should address persistent confounding and uncertainty quantification.

Acknowledgments and Disclosure of Funding

References

Adamy Jürgen. Nonlinear Systems and Controls. Berlin, Heidelberg: Springer Berlin Heidelberg, 2024. 2. 2nd ed. 2024.

Bang Heejung, Robins James M. Doubly Robust Estimation in Missing Data and Causal Inference Models // Biometrics. XII 2005. 61, 4. 962–973.

Bernard Pauline. Observer Design for Nonlinear Systems. 479. Cham: Springer International Publishing, 2019. (Lecture Notes in Control and Information Sciences).

Bica Ioana, Alaa Ahmed M., Jordon James, Schaar Mihaela van der. Estimating Counterfactual Treatment Outcomes over Time Through Adversarially Balanced Representations. II 2020.

Bongers Stephan, Blom Tineke, Mooij Joris M. Causal Modeling of Dynamical Systems // arXiv.org preprint. III 2022. arXiv:1803.08784v4 [cs.AI].

Brouwer Edward De, Hernández Javier González, Hyland Stephanie. Predicting the impact of treatments over time with uncertainty aware neural differential equations. II 2022.

Chen Ricky T. Q., Rubanova Yulia, Bettencourt Jesse, Duvenaud David K. Neural Ordinary Differential Equations // Advances in Neural Information Processing Systems. 31. 2018.

Geng Changran, Paganetti Harald, Grassberger Clemens. Prediction of Treatment Response for Combined Chemo- and Radiation Therapy for Non-Small Cell Lung Cancer Patients Using a Bio-Mathematical Model // Scientific Reports. X 2017. 7, 1. 13542.

Goldberger Ary L., Amaral Luis A. N., Glass Leon, Hausdorff Jeffrey M., Ivanov Plamen Ch, Mark Roger G., Mietus Joseph E., Moody George B., Peng Chung-Kang, Stanley H. Eugene. PhysioBank, PhysioToolkit, and PhysioNet // Circulation. VI 2000.

Hensman James, Durrande Nicolas, Solin Arno. Variational Fourier Features for Gaussian Processes // Journal of Machine Learning Research. 2018. 18, 151. 1–52.

Hernán Miguel, Robins James M. Causal inference: what if. Boca Raton: Taylor and Francis, 2023. First edition.

Hess Konstantin, Feuerriegel Stefan. Stabilized Neural Prediction of Potential Outcomes in Continuous Time. 2025.

Hess Konstantin, Frauen Dennis, Melnychuk Valentyn, Feuerriegel Stefan. IGC-Net for conditional average potential outcome estimation over time // The Fourteenth International Conference on Learning Representations. 2026.

Hofford Mackenzie R, Yu Sean C, Johnson Alistair E W, Lai Albert M, Payne Philip R O, Michelson Andrew P. OpenSep: a generalizable open source pipeline for SOFA score calculation and Sepsis-3 classification // JAMIA Open. XII 2022. 5, 4. ooac105.

Jazwinski Andrew H. Stochastic processes and filtering theory. Mineola, N.Y: Dover Publications, 2007. Dover ed. (Dover books on engineering).

Johnson Alistair E. W., Bulgarelli Lucas, Shen Lu, Gayles Alvin, Shammout Ayad, Horng Steven, Pollard Tom J., Hao Sicheng, Moody Benjamin, Gow Brian, Lehman Li-wei H., Celi Leo A., Mark Roger G. MIMIC-IV, a freely accessible electronic health record dataset // Scientific Data. I 2023. 10, 1. 1.

Johnson Alistair E. W., Pollard Tom J., Shen Lu, Lehman Li-Wei H., Feng Mengling, Ghassemi Mohammad, Moody Benjamin, Szolovits Peter, Celi Leo Anthony, Mark Roger G. MIMIC-III, a freely accessible critical care database // Scientific Data. V 2016. 3. 160035.

Jong Johann de, Emon Mohammad Asif, Wu Ping, Karki Reagon, Sood Meemansa, Godard Patrice, Ahmad Ashar, Vrooman Henri, Hofmann-Apitius Martin, Fröhlich Holger. Deep learning for clustering of multivariate clinical patient trajectories with missing values // GigaScience. XI 2019. 8, 11. giz134.

Kallenberg Olav. Foundations of Modern Probability. 99. 2021. 3. (Probability Theory and Stochastic Modelling).

Kidger Patrick, Morrill James, Foster James, Lyons Terry. Neural Controlled Differential Equations for Irregular Time Series. XI 2020.

Li Rui, Shahn Zach, Li Jun, Lu Mingyu, Chakraborty Prithwish, Sow Daby, Ghalwash Mohamed, Lehman Li-wei H. G-Net: A Deep Learning Approach to G-computation for Counterfactual Outcome Prediction Under Dynamic Treatment Regimes. III 2020.

Lok Judith J. Statistical modeling of causal effects in continuous time // Ann. Statist. VI 2008. 36, 3.

Melnychuk Valentyn, Frauen Dennis, Feuerriegel Stefan. Causal Transformer for Estimating Counterfactual Outcomes. VI 2022.

Murphy S. A. Optimal Dynamic Treatment Regimes // Journal of the Royal Statistical Society Series B: Statistical Methodology. IV 2003. 65, 2. 331–355. _eprint: https://academic.oup.com/jrsssb/article-pdf/65/2/331/49684411/jrsssb_65_2_331.pdf.

Pearl Judea. Causality: models, reasoning, and inference. Cambridge, U.K. ; New York: Cambridge University Press, 2000.

Peters Jonas, Janzing Dominik, Schölkopf Bernhard. Elements of causal inference: foundations and learning algorithms. Cambridge, Mass: The MIT press, 2017. (Adaptive computation and machine learning).

Robins James. 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. I 1986. 7, 9. 1393–1512.

Robins James M. Optimal Structural Nested Models for Optimal Sequential Decisions // Proceedings of the Second Seattle Symposium in Biostatistics: Analysis of Correlated Data. New York, NY: Springer New York, 2004. 189–326.

Rubanova Yulia, Chen Ricky T. Q., Duvenaud David. Latent ODEs for Irregularly-Sampled Time Series. 2019.

Rubin Donald B. Randomization Analysis of Experimental Data: The Fisher Randomization Test Comment // Journal of the American Statistical Association. IX 1980. 75, 371. 591.

Rytgaard Helene C., Gerds Thomas A., Van Der Laan Mark J. Continuous-time targeted minimum loss-based estimation of intervention-specific mean outcomes // Ann. Statist. X 2022. 50, 5.

Schulam Peter, Saria Suchi. Reliable Decision Support using Counterfactual Models. II 2018.

Seedat Nabeel, Imrie Fergus, Bellot Alexis, Qian Zhaozhi, Schaar Mihaela van der. Continuous-Time Modeling of Counterfactual Outcomes Using Neural Controlled Differential Equations. 2022.

Thoral Patrick J., Peppink Jan M., Driessen Ronald H., Sijbrands Eric J. G., Kompanje Erwin J. O., Kaplan Lewis, Bailey Heatherlee, Kesecioglu Jozef, Cecconi Maurizio, Churpek Matthew, Clermont Gilles, Van Der Schaar Mihaela, Ercole Ari, Girbes Armand R. J., Elbers Paul W. G. Sharing ICU Patient Data Responsibly Under the Society of Critical Care Medicine/European Society of Intensive Care Medicine Joint Data Science Collaboration: The Amsterdam University Medical Centers Database (AmsterdamUMCdb) Example* // Critical Care Medicine. VI 2021. 49, 6. e563–e577.

Van Der Laan Mark J., Gruber Susan. Targeted Minimum Loss Based Estimation of Causal Effects of Multiple Time Point Interventions // The International Journal of Biostatistics. V 2012. 8, 1.

Van Handel Ramon. Observability and nonlinear filtering // Probability Theory and Related Fields. IX 2009. 145, 1-2. 35–74.

Vincent J. L., Moreno R., Takala J., Willatts S., De Mendonça A., Bruining H., Reinhart C. K., Suter P. M., Thijs L. G. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine // Intensive Care Medicine. VII 1996. 22, 7. 707–710.

Wendland Philipp, Birkenbihl Colin, Gomez-Freixa Marc, Sood Meemansa, Kschischo Maik, Fröhlich Holger. Generation of realistic synthetic data using Multimodal Neural Ordinary Differential Equations // npj Digital Medicine. VIII 2022. 5, 1. 122.

Wendland Philipp, Kschischo Maik. Dynamic Dose Optimization Under Side Effect Constraints Using Counterfactual Outcome Prediction // Artificial Intelligence in Medicine. 15735. Cham: Springer Nature Switzerland, 2025. 428–433. Series Title: Lecture Notes in Computer Science.

Wendland Philipp, Schenkel-Häger Christof, Wenningmann Ingobert, Kschischo Maik. An optimal antibiotic selection framework for Sepsis patients using Artificial Intelligence // NPJ digital medicine. XI 2024. 7, 1. 343.

Xu Ziqi, Cheng Debo, Li Jiuyong, Liu Jixue, Liu Lin, Yu Kui. Causal Inference with Conditional Front-Door Adjustment and Identifiable Variational Autoencoder // The Twelfth International Conference on Learning Representations. 2024.

A Technical Appendix and Supplementary Material

A.1 Derivation of equation 4

In the main text we derived only a discrete version of the adjustment formula (4) and here we provide the continuous time version. Assume we have observed a treatment and outcome history $\boldsymbol{o}_{[0,t)} = \{\boldsymbol{a}_{[0,t)}, \boldsymbol{y}_{[0,t)}\}$ before time t and also the outcome $\boldsymbol{Y}_t = \boldsymbol{y}_t$ at time t. We want to compute the probability density $P_{\boldsymbol{Y}_{t+s}(\boldsymbol{a}_{[t,t+s)})|\boldsymbol{Y}_t,\boldsymbol{O}_{[0,t)}}(\boldsymbol{y}_{t+s}|\boldsymbol{y}_t, \boldsymbol{o}_{[0,t)})$ for the potential outcome $\boldsymbol{Y}_{t+s}\left(\boldsymbol{a}_{[t,t+s)}\right)$ under the hypothetical intervention path $\boldsymbol{a}_{[t,t+s)}$ applied in the time interval $[t, t+s)$.

Let us partition the data assimilation time interval $[0, t)$ into K subintervals $[t_k, t_{k+1})$ with $t_0 = 0, t_K = t$ and $[0, t) = \cup_{k=1}^{K}[t_k, t_{k+1})$. Similarly, partition the prediction interval $[t, t+s)$ into L subintervals with $[t_l, t_{l+1})$ with $t_{K+1} = t, t_{K+L} = t+s$ and $[t, t+s) = \cup_{l=K+1}^{K+L}[t_l, t_{l+1})$.

For each $t_k \in [0, t)$ and $t_l \in [t, t+s)$ we identify the variables in the CFD (2) with variables in our model as

$$A \leftarrow\!\!\prec \boldsymbol{A}_t, \quad Y \leftarrow\!\!\prec \boldsymbol{Y}_{t_l}, \quad M \leftarrow\!\!\prec \boldsymbol{Z}_{t_l}, \quad W \leftarrow\!\!\prec \{\boldsymbol{Z}_t, \boldsymbol{Y}_t, \boldsymbol{A}_{t_k}, \boldsymbol{Y}_{t_k}\} = \{\boldsymbol{Z}_t, \boldsymbol{Y}_t, \boldsymbol{O}_{t_k}\},$$

with $\boldsymbol{O}_{t_k} = \{\boldsymbol{A}_{t_k}, \boldsymbol{Y}_{t_k}\}$. We can check graphically (replace t_k by 0 and t_l by $t+s$ in Fig. 1) that $M = \boldsymbol{Z}_{t_l}$ fulfills the CFD 3.2 with respect to $(A, Y) = (\boldsymbol{A}_t, \boldsymbol{Y}_{t_l})$ with the conditioning variables $W = \{\boldsymbol{Z}_t, \boldsymbol{Y}_t, \boldsymbol{O}_{t_k}\}$.

In order to apply the CFD to sets of variables we define

$$\boldsymbol{A}_{0:K} := \{\boldsymbol{A}_{t_k} | t_k \in [0, t)\}, \qquad \boldsymbol{Y}_{0:K} := \{\boldsymbol{Y}_{t_k} | t_k \in [0, t)\}, \qquad \boldsymbol{A}_{K+1:L} := \{\boldsymbol{A}_{t_l} | t_l \in [t, t+s)\}.$$

Selecting the following sets in Theorem 3.2

$$A \leftarrow\!\!\prec \boldsymbol{A}_{K+1:L}, \quad Y \leftarrow\!\!\prec \boldsymbol{Y}_{t+s}, \quad M \leftarrow\!\!\prec \boldsymbol{Z}_{t+s}, \quad W \leftarrow\!\!\prec \{\boldsymbol{Z}_t, \boldsymbol{Y}_t, \boldsymbol{A}_{0:K}, \boldsymbol{Y}_{0:K}\} = \{\boldsymbol{Z}_t, \boldsymbol{Y}_t, \boldsymbol{O}_{0:K}\},$$

we see that $M = \boldsymbol{Z}_{t+s}$ fulfills the CFD relative to $(A, Y) = (\boldsymbol{A}_{K+1:L}, \boldsymbol{Y}_{t+s})$ with the conditioning set $W = \{\boldsymbol{Z}_t, \boldsymbol{Y}_t, \boldsymbol{O}_{0:K}\}$. Applying the adjustment formula of the CFD (2) we have for the probability density of the potential outcome $\boldsymbol{Y}_{t+s}$ under the treatment sequence $\boldsymbol{a}_{K+1:L}$

$$P_{\boldsymbol{Y}_{t+s}(\boldsymbol{A}_{K+1:L})}(\boldsymbol{y}_{t+s}) = \int_{\boldsymbol{z}_{t+s}} \int_{\boldsymbol{a}'_{K+1:L}} \int_{\boldsymbol{z}_t, \boldsymbol{y}_t, \boldsymbol{o}_{0:K}} \big\{ p(\boldsymbol{y}_{t+s} | \boldsymbol{a}'_{K+1:L}, \boldsymbol{z}_{t+s}, \boldsymbol{z}_t, \boldsymbol{y}_t, \boldsymbol{o}_{0:K}) \tag{8}$$

$$\times\, p(\boldsymbol{a}'_{K+1:L} | \boldsymbol{z}_t, \boldsymbol{y}_t, \boldsymbol{o}_{0:K})\, p(\boldsymbol{z}_{t+s} | \boldsymbol{a}_{K+1:L}, \boldsymbol{z}_t, \boldsymbol{y}_t, \boldsymbol{o}_{0:K})\, p(\boldsymbol{z}_t, \boldsymbol{y}_t, \boldsymbol{o}_{0:K}) \big\}. \tag{9}$$

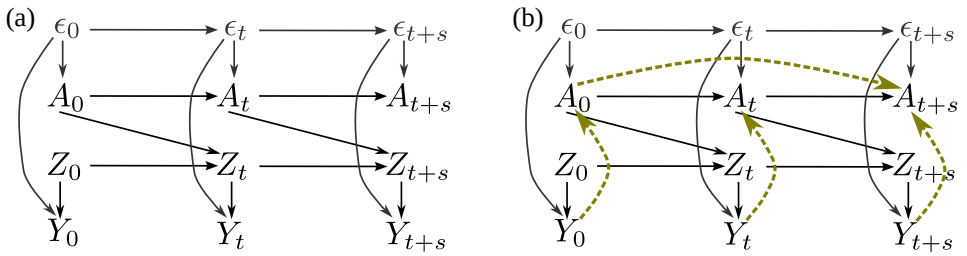


Figure 5: (a) Temporal DAG for treatment A, observation Y, and latent state Z at three consecutive time points. In contrast to Fig. 1, the unobserved process ϵ jointly affecting treatment assignment and observations exhibits temporal dependence. (b) Under a dynamic treatment regime, treatment additionally depends on the observed treatment–outcome history, as indicated by the green dashed arrows.

Now we use the independence relations from assumptions 4.2 and 4.1

$$P(\boldsymbol{y}_{t+s}|\boldsymbol{a}'_{K+1:L}, \boldsymbol{z}_{t+s}, \boldsymbol{z}_t, \boldsymbol{y}_t, \boldsymbol{o}_{0:K}) = p(\boldsymbol{y}_{t+s}|\boldsymbol{z}_{t+s})$$
$$p(\boldsymbol{z}_{t+s}|\boldsymbol{a}_{K+1:L}, \boldsymbol{z}_t, \boldsymbol{y}_t, \boldsymbol{o}_{0:K}) = p(\boldsymbol{z}_{t+s}|\boldsymbol{a}_{K+1:L}, \boldsymbol{z}_t)$$

and find for the probability density of the potential outcome

$$\begin{aligned} P_{\boldsymbol{Y}_{t+s}(\boldsymbol{A}_{K+1:L})}(\boldsymbol{y}_{t+s}) &= \int_{\boldsymbol{z}_{t+s}} \int_{\boldsymbol{a}'_{K+1:L}} \int_{\boldsymbol{z}_t, \boldsymbol{y}_t, \boldsymbol{o}_{0:K}} \{p(\boldsymbol{y}_{t+s}|\boldsymbol{z}_{t+s}) \\ &\quad \times p(\boldsymbol{a}'_{K+1:L}|\boldsymbol{z}_t, \boldsymbol{y}_t, \boldsymbol{o}_{0:K})\, p(\boldsymbol{z}_{t+s}|\boldsymbol{a}_{K+1:L}, \boldsymbol{z}_t)\, p(\boldsymbol{z}_t, \boldsymbol{y}_t, \boldsymbol{o}_{0:K})\} \\ &= \int_{\boldsymbol{z}_{t+s}} \int_{\boldsymbol{z}_t, \boldsymbol{y}_t, \boldsymbol{o}_{0:K}} p(\boldsymbol{y}_{t+s}|\boldsymbol{z}_{t+s}) p(\boldsymbol{z}_{t+s}|\boldsymbol{a}_{K+1:L}, \boldsymbol{z}_t)\, p(\boldsymbol{z}_t|\boldsymbol{y}_t, \boldsymbol{o}_{0:K})\, p(\boldsymbol{y}_t, \boldsymbol{o}_{0:K}). \end{aligned}$$

From this, we read

$$P_{\boldsymbol{Y}_{t+s}(\boldsymbol{a}_{K+1:L})|\boldsymbol{Y}_t, \boldsymbol{O}_{0:K}}(\boldsymbol{y}_{t+s}|\boldsymbol{y}_t, \boldsymbol{o}_{0:K}) \tag{10}$$
$$= \int_{\boldsymbol{z}_{t+s}} \int_{\boldsymbol{z}_t} p(\boldsymbol{y}_{t+s}|\boldsymbol{z}_{t+s}) p(\boldsymbol{z}_{t+s}|\boldsymbol{a}_{K+1:L}, \boldsymbol{z}_t)\, p(\boldsymbol{z}_t|\boldsymbol{y}_t, \boldsymbol{o}_{0:K}). \tag{11}$$

This formula is valid for any partitions of the data assimilation interval $[0, t)$ and the forecast interval $[t, t+s)$. Making the partitions finer and finer by sending $K \to \infty$ and $L \to \infty$ allows us to replace $\boldsymbol{o}_{0:K}$ and by $\boldsymbol{o}_{[0,t)}$ and $\boldsymbol{a}_{K+1:L}$ by $\boldsymbol{a}_{[t,t+s)}$ and we obtain (4). The model forecast is deterministic in (3) and we have $p(\boldsymbol{z}_{t+s}|\boldsymbol{a}_{[t,t+s)}, \boldsymbol{z}_t) = \delta\left(\boldsymbol{z}_{t+s} - \boldsymbol{z}_t - \int_t^{t+s} \boldsymbol{f}(\boldsymbol{z}_\tau, \boldsymbol{a}_\tau)\, d\tau\right)$.

Applying the continuum limit in (10) only to the hypothetical treatment trajectory ($L \to \infty$) we have

$$\begin{aligned} &P_{\boldsymbol{Y}_{t+s}(\boldsymbol{a}_{[t,t+s)})|\boldsymbol{Y}_t, \boldsymbol{O}_{0:K}}(\boldsymbol{y}_{t+s}|\boldsymbol{y}_t, \boldsymbol{o}_{0:K}) \\ &\quad = \int_{\boldsymbol{z}_{t+s}} \int_{\boldsymbol{z}_t} p(\boldsymbol{y}_{t+s}|\boldsymbol{z}_{t+s}) p(\boldsymbol{z}_{t+s}|\boldsymbol{a}_{[t,t+s)}, \boldsymbol{z}_t)\, p(\boldsymbol{z}_t|\boldsymbol{y}_t, \boldsymbol{o}_{0:K}). \end{aligned}$$

If the discretization time points $t_k \in [0, t)$ for the data assimilation interval $[0, t)$ coincide with the time points for the observed data, we have an adjustment formula for discrete time observations and continuous time treatment trajectories. Here, $p(\boldsymbol{z}_t|\boldsymbol{y}_t, \boldsymbol{o}_{0:K})$ is the filtering distribution for the continuous time ODE model 3 with discrete time observations (Jazwinski, 2007).

A.2 An adjustment formula for serially correlated hidden confounders

Assumption 4.2 is compatible with temporally independent unobserved disturbances $\boldsymbol{\epsilon}_t$ that affect both treatment assignment and contemporaneous observations, as shown in Fig. 1. If $\boldsymbol{\epsilon}_t$ is serially correlated, as illustrated in Fig. 5, the observation at time $t+s$ may remain statistically dependent on the treatment trajectory even after conditioning on the latent state $\boldsymbol{Z}_{t+s}$. Consequently, the conditional independence

$$p(\boldsymbol{y}_{t+s}|\boldsymbol{a}'_{K+1:L}, \boldsymbol{z}_{t+s}, \boldsymbol{z}_t, \boldsymbol{y}_t, \boldsymbol{o}_{0:K}) = p(\boldsymbol{y}_{t+s}|\boldsymbol{z}_{t+s})$$

used in the preceding derivation no longer holds.

Nevertheless, provided that the CFD criterion remains satisfied, the potential-outcome distribution is still identified by the unsimplified adjustment formula 8. For a finite partition, conditioning the CFD adjustment formula on the observed history gives

$$\begin{aligned} &P_{\boldsymbol{Y}_{t+s}(\boldsymbol{a}_{K+1:L})|\boldsymbol{Y}_t,\boldsymbol{O}_{0:K}}\left(\boldsymbol{y}_{t+s}|\boldsymbol{y}_t,\boldsymbol{o}_{0:K}\right) \\ &\quad= \int_{\boldsymbol{z}_t}\int_{\boldsymbol{z}_{t+s}}\int_{\boldsymbol{a}'_{K+1:L}} \{p\left(\boldsymbol{y}_{t+s}|\boldsymbol{a}'_{K+1:L},\boldsymbol{z}_{t+s},\boldsymbol{z}_t,\boldsymbol{y}_t,\boldsymbol{o}_{0:K}\right) \\ &\qquad\quad\times p\left(\boldsymbol{a}'_{K+1:L}|\boldsymbol{z}_t,\boldsymbol{y}_t,\boldsymbol{o}_{0:K}\right) p\left(\boldsymbol{z}_{t+s}|\boldsymbol{a}_{K+1:L},\boldsymbol{z}_t\right) p\left(\boldsymbol{z}_t|\boldsymbol{y}_t,\boldsymbol{o}_{0:K}\right)\}. \end{aligned} \tag{12}$$

Here, $\boldsymbol{a}_{K+1:L}$ denotes the hypothetical treatment sequence of interest, whereas $\boldsymbol{a}'_{K+1:L}$ is integrated with respect to its observational distribution.

In the continuum limit, $\boldsymbol{a}'_{K+1:L}$ becomes an alternative treatment trajectory $\boldsymbol{a}'_{[t,t+s)}$. Because the space of continuous-time trajectories is generally infinite-dimensional, the corresponding integration is expressed using the conditional path measure (Kallenberg, 2021)

$$P\left(d\boldsymbol{a}'_{[t,t+s)}|\boldsymbol{z}_t,\boldsymbol{y}_t,\boldsymbol{o}_{[0,t)}\right).$$

The continuous-time adjustment formula is therefore

$$\begin{aligned} &P_{\boldsymbol{Y}_{t+s}(\boldsymbol{a}_{[t,t+s)})|\boldsymbol{Y}_t,\boldsymbol{O}_{[0,t)}}\left(\boldsymbol{y}_{t+s}|\boldsymbol{y}_t,\boldsymbol{o}_{[0,t)}\right) \\ &\quad= \int_{\boldsymbol{z}_t}\int_{\boldsymbol{z}_{t+s}}\int_{\mathcal{A}_{[t,t+s)}} \Big\{p\left(\boldsymbol{y}_{t+s}|\boldsymbol{a}'_{[t,t+s)},\boldsymbol{z}_{t+s},\boldsymbol{z}_t,\boldsymbol{y}_t,\boldsymbol{o}_{[0,t)}\right) \\ &\qquad\quad\times\ P\left(d\boldsymbol{a}'_{[t,t+s)}|\boldsymbol{z}_t,\boldsymbol{y}_t,\boldsymbol{o}_{[0,t)}\right) p\left(\boldsymbol{z}_{t+s}|\boldsymbol{a}_{[t,t+s)},\boldsymbol{z}_t\right) p\left(\boldsymbol{z}_t|\boldsymbol{y}_t,\boldsymbol{o}_{[0,t)}\right)\Big\} \end{aligned} \tag{13}$$

where $\mathcal{A}_{[t,t+s)}$ denotes the space of admissible treatment trajectories. Thus, observability still permits identification through the latent mediator, but serial dependence of the hidden confounder prevents the observational treatment-path distribution from being integrated out.

Equation 13 could in principle be approximated by learning the conditional observational treatment process, sampling alternative treatment trajectories from this process, and averaging the conditional outcome density over these trajectories. Such an implementation additionally requires estimation of the complete treatment-path distribution and suitable path-space positivity. However, its non-parametric estimation additionally requires path-space and mediator positivity, which may fail for deterministic latent dynamics. Practical implementations may therefore require stochastic latent dynamics or an augmented observable state that captures the temporally persistent confounding process.

The ObsNODE implementation considered in this work instead uses Assumption 4.2, under which the outcome density is independent of $\boldsymbol{a}'_{[t,t+s)}$ given $\boldsymbol{Z}_{t+s}$. The path integral then equals one, and (13) reduces to the tractable adjustment formula (4).

A.3 Details on training the ObsNODE model

The datasets were split into training, validation, and test sets and feature-wise Z-score normalization was applied using statistics computed on the training set only. The normalized dataset is subsequently used for model training.

Each multivariate time series was segmented at a decision time point t_c into two disjoint subsequences. The prefix segment consists of all observations prior to t_c and is used to condition the recognition model. The model infers an initial state $\boldsymbol{z}_0$ for each individual trajectory using an an LSTM-based recognition network (Rubanova et al., 2019).

To handle missing observations, the recognition LSTM is augmented with an explicit imputation mechanism. The indicator function

$$\mathcal{I}_{tij} = \begin{cases} 1: & y_{tij} \text{ is observed} \\ 0: & y_{tij} \text{ is missing.} \end{cases}$$

indicates whether the data for the $j-$th component of the $i-$th unit at time t is observed or not, The imputation layer is defined as the transformation

$$y_{tij} = y_{tij} \cdot \mathcal{I}_{tij} + b_j \cdot (1 - \mathcal{I}_{tij}),$$

where b_j are learnable parameters (Jong de et al., 2019; Wendland et al., 2022). This formulation replaces missing entries with trainable values that are optimized jointly with the LSTM parameters in an end-to-end fashion.

The latent initial states $\boldsymbol{z}_0$ inferred by the LSTM, together with the control function $\boldsymbol{a}_t$ are subsequently passed as inputs to the Neural Differential Equation (NODE) model. The observations after the decision time point t_c are used for loss computation and subsequent optimization. Model parameters are optimized using the Adam optimizer.

The ObsNODE model supports both direct long-horizon forecasting and a recursive scheme, similar in spirit to ODE-RNN (Rubanova et al., 2019). In the long-horizon setting, the state estimated by the LSTM recognition model is used to generate a forecast over the entire prediction window. In the recursive setting, forecasts are produced iteratively: at each step, the ODE model generates a short-term prediction, which is appended to the observation and treatment history sequence observed up to t_c. Conditioned on this augmented sequence, the LSTM updates the latent state, which then serves as the initial condition for the next ODE rollout. This process repeats until the full horizon is covered, without using any ground-truth observations beyond t_c. We find empirically that augmenting the sequence with intermediate predictions can improve stability for longer forecasting horizons in some settings.

A.3.1 Code availability

ObsNODE is implemented in PyTorch. Code is available at `https://github.com/JenniferJaschob/ObsNODE.git`. The repository includes scripts for data preprocessing, simulation, model training, and evaluation, together with configuration files specifying all hyperparameters used in the experiments.

A.4 Details for Computational Experiments

A.4.1 Details for Synthetic Data Simulation

The synthetic data set is simulated from a dynamic model for tumor volume $V(t)$ (in cm^3) (Geng et al., 2017; Bica et al., 2020) and net body weight $W^{net}(t)$ (in kg) in response to a dynamic dosing trajectory for chemotherapy $C(t)$ and radiotherapy $d(t)$ combination treatment (Wendland, Kschischo, 2025):

$$\frac{dV(t)}{dt} = \left(\rho \log\left(\frac{K}{V(t)}\right) - \beta_c C(t) - \left(\alpha_r d(t) + \beta_r d(t)^2\right) + \epsilon(t)\right) V(t) \tag{14a}$$

$$\frac{dW^{net}(t)}{dt} = k_g W^{net}(t)\left(1 - \frac{W^{net}(t)}{G}\right) - \beta_{wc} C(t) - \alpha_{wr} d(t) - \lambda V(t). \tag{14b}$$

The tumor volume is assumed to follow a Gompertz growth function, which is inhibited by the chemotherapy and radiotherapy dosing. The Gaussian zero mean white noise process $\epsilon(t)$ renders the tumour volume a stochastic variable.

Net body weight $W^{net}(t)$ is modeled as a logistic growth process, modified by tumor volume and therapy-induced weight loss (Wendland, Kschischo, 2025). Net body weight $W^{net}(t) = W(t) - \rho_d V(t)$ is defined as body weight $W(t)$ minus the tumor weight, with ρ_d denoting the tumor density. The parameter G corresponds to the saturation level, which in this case is set to the initial weight. The parameters β_{wc}, α_{wr}, and λ determine the weight loss contributions of chemotherapy, radiotherapy, and disease impact, respectively.

Treatment doses are modeled as time dependent random variables

$$C(t) = C_{max}\, \sigma\left(\frac{\gamma_c\, \alpha_c^{dose}}{D_{max}}(\overline{D}(t) - \delta_c)\right) \tag{15a}$$

$$d(t) = d_{max}\, \sigma\left(\frac{\gamma_r\, \alpha_r^{dose}}{D_{max}}(\overline{D}(t) - \delta_r)\right), \tag{15b}$$

where σ is the softmax function and $C_{max} = 14\,\text{mg/m}^3$ and $d_{max} = 3\,\text{Gy}$ are the maximum doses for for chemo- and radiotherapy, respectively. The prescribed treatment doses depend on the outcome via the mean tumor diameter $\overline{D}(t)$ over the last 15 days (Melnychuk et al., 2022). The parameter $D_{max} = 13\,cm$ is the maximum tumor diameter and we set $\delta_c = \delta_r = \frac{D_{max}}{2}$. The degree of time dependent confounding is given by the parameters γ_c and γ_r, with range $\gamma = \gamma_c = \gamma_r \in [1, 8]$. The values of α_c^{dose} and α_c^{dose} are sampled uniformly in the range from 1cm to 4cm to simulate variability in individual dosing decisions. The IGC-Net model can only process binary treatments indicators $\boldsymbol{a}_t$ and we interprete $C(t)/C_{max}$ and $d(t)/d_{max}$ for IGC-Net as probabilities for the Bernoulli-variables of treatment for chemo- and radiotherapy, respectively.

We simulate 3,000 patients and use 1,000 patients each for training, validation, and testing. Chemotherapy dose $C(t)$ and radiotherapy dose $d(t)$ was adjusted in monthly treatment cycles by sampling from (15). To model patient variability, we sampled patient specific parameters from Gaussian distributions, see Table 6.

MODEL	VARIABLE	PARAMETER	DISTRIBUTION	DISTRIBUTION VALUE (μ, σ)
TUMOR	GROWTH PARAMETER	ρ	NORMAL	$7 \times 10^{-5},\ 0.00723$
TUMOR	CARRYING CAPACITY	K	CONSTANT	30
TUMOR	RADIO CELL KILL (α)	α_r	NORMAL	$0.0398,\ 0.168$
TUMOR	RADIO CELL KILL (β)	β_r	-	SET $\alpha/\beta = 10$
TUMOR	CHEMO CELL KILL	β_c	NORMAL	$0.028,\ 0.0007$
TUMOR	NOISE	ϵ_{vt}	NORMAL	$0,\ 0.01$
WEIGHT	GROWTH PARAMETER	ρ_w	NORMAL	$14 \times 10^{-5},\ 10^{-5}$
WEIGHT	CARRYING CAPACITY	K_w	CONSTANT	INITIAL WEIGHT
WEIGHT	RADIO INFLUENCE	α_{wr}	NORMAL	$0.004125,\ 10^{-4}$
WEIGHT	CHEMO INFLUENCE	β_{wc}	NORMAL	$0.001775,\ 2 \times 10^{-4}$
WEIGHT	CANCER INFLUENCE	λ	NORMAL	$31 \times 10^{-5},\ 15 \times 10^{-6}$
WEIGHT	NOISE	ϵ_{wt}	NORMAL	$0,\ 0.0015$

Table 6: Parameter of the cancer growth simulation model (Wendland, Kschischo, 2025).

We generated five counterfactual treatment scenarios (CF_1–CF_5). For CF_1 and CF_2 , we sampled two different treatment sequences by independently and uniformly selecting one of four predefined treatment application levels at each time step (chemotherapy: {0, 3, 5, 7}, radiotherapy: {0, 1, 2, 3}). These sampled dose sequences were then fixed and applied deterministically to all patients after the intervention time. These values represent the applied values, while the effective dose is modeled via the pharmacokinetic drug decay process. CF_1 and CF_2 therefore represent randomly assigned alternative treatment sequences. CF_3 denotes a strategy consisting of maximum radiotherapy without chemotherapy, while CF_4 denotes a strategy consisting of maximum chemotherapy without radiotherapy. CF_5 corresponds to no treatment.

A.4.2 Details for Semi-Synthetic Data Simulation

To generate semi-synthetic data (Schulam, Saria, 2018) we have modified the approach from (Melnychuk et al., 2022). We use the sepsis patient cohort (3557 patients) from the MIMIC-IV dataset (Johnson et al., 2023) (see also A.4.3) and select three vital signs as time dependent hidden confounders $\boldsymbol{\epsilon}_t$. Then, we simulate untreated outcomes $j \in \{1, \ldots, d_y\}$ as

$$\tilde{Y}_t^{j,(i)} = \alpha_s^j \textit{ B-spline(t)} + \alpha_g^j\, g^{j,(i)}(t) + \alpha_\phi^j\, \phi_Y^j(\boldsymbol{\epsilon}_t^{(i)}) + \eta_t^{j,(i)} \tag{16}$$

with noise $\eta_t^{j,(i)} \sim \mathcal{N}(0,\ 0.005^2)$ and weight coefficients α_s^j, α_g^j and α_ϕ^j. The term *B-spline*(t) represents a mixture of three cubic B-splines. The function $g^{j,(i)}$ is independently sampled for each patient i from a Gaussian process with a Matérn kernel, and ϕ_Y^j is sampled via random Fourier features (RFF) (Hensman et al., 2018).

Synthetic binary treatments $\boldsymbol{A}_t = (A_t^1, \ldots, A_t^{d_a}) \in \{0, 1\}^{d_a}$ with confounding are simulated via

$$A_t^l \sim \text{Bernoulli}(P_{A_t^l}), \qquad P_{A_t^l} = \sigma\left(\gamma_A^l\, \bar{Y}_{t-1}^{A,l} + \gamma_\epsilon^l \Phi_\epsilon^l(\boldsymbol{\epsilon}_t) + b_l\right),$$

where σ is the sigmoid function. The treatment probability $P_{A_t^l}$ for the $l-$th treatment component A_t^l depends on the outcome history, represented by an average over a subset of the previous treated outcomes $\bar{Y}_{t-1}^{A,l}$ (Melnychuk et al., 2022). Hidden confounding of the treatment is represented via the function Φ_ϵ^l sampled from a RFF approximation of a Gaussian process. The parameters γ_A^l and γ_ϵ^l control the confounding strength and b_l is a bias parameter.

The time dependent treatment effect

$$E_t^j = \sum_{k=t-w^l}^{t} \frac{\min_{l=1,\dots,d_a} \mathbf{1}_{[A_k^l=1]} P_{A_k^l} \beta_{lj}}{(w^l - k)^2}. \tag{17}$$

is bounded by the maximum effect size $\beta_{lj} \in \{0, \beta\}$ of treatment component l on outcome j. It is either zero for outcomes j not affected by treatment l or equals the constant β. Finally, the treated outcomes are obtained from the untreated outcomes (16) by adding the treatment effects

$$Y_t^{j,(i)} = \tilde{Y}_t^{j,(i)} + E_t^j \,. \tag{18}$$

Parameters for the semi-synthetic sepsis data are given in Table 7.

Variable	**Parameter**	**Value**
Outcome	α	2
Outcome	α_g	0.5
Outcome	α_e	1
Outcome	ν	20
Outcome	Exogeneous Variables	Creatinine, Bilirubin, ALT
Treatment 1; 2	γ_z	0.3; 0.3
Treatment 1; 2	γ_0	0.3; 0.1
Treatment 1; 2	bias	-2; -2
Treatment 1; 2	ν	20; 20
Treatment 1; 2	Confounding Variables	Creatinine; Bilirubin, ALT

Table 7: **Parameters of the data-simulation model for semi-synthetic data (Wendland, Kschischo, 2025).**

We generated five counterfactual treatment scenarios CF_k for $k \in 1, \dots, 5$ by modifying the treatment assignment mechanism during simulation. CF_1 corresponds to a randomized treatment policy, with treatment probabilities sampled randomly. For CF_2–CF_5, treatment probabilities are deterministically set to one or zero, corresponding to maximal or minimal treatment, respectively. CF_2 implements a strategy with maximum first treatment and no subsequent treatment, whereas CF_3 represents a strategy with no first treatment and maximum second treatment. CF_4 corresponds to a no-treatment strategy, while CF_5 represents a strategy with maximum treatment for all time points after the decision time point t_c.

A.4.3 Details on Experiments with Real-World Data

MIMICIV sepsis cohort: We use the MIMIC-IV electronic health record dataset collected at Beth Israel Deaconess Medical Center, which contains data from approximately 450,000 hospital admissions (Johnson et al., 2023). The dataset includes demographic information, laboratory values, vital signs, medications, and diagnoses.

For our analysis we focus on sepsis patients identified using the OpenSep pipeline according to the Sepsis-3 criteria (Hofford et al., 2022). Sepsis cases were defined by an increase in the Sequential Organ Failure Assessment (SOFA) score of at least two points within 48 hours before or 24 hours after a suspected infection.

The SOFA score is used as the clinical outcome variable. It summarizes dysfunction in six organ systems (respiratory, cardiovascular, central nervous system, coagulation, liver, and kidney), each scored from 0 to 4 according to the severity of dysfunction (Vincent et al., 1996). An overview of the subscores is given in Table 8.

ORGAN	PARAMETER	UNIT	1	2	3	4
RESPIRATION	PaO_2/FiO_2	$mmHg$	<400	<300	<200	<100
COAGULATION	PLATELETS	μl	<150.000	<100.00	<50.000	<20.000
LIVER	BILIRUBIN	MG/DL	1.2-1.9	2.0-5.9	6.0-11.9	>12.0
CARDIO-VASCULAR	MAP OR VASOPRESSORS	MMHG OR $\mu g/kg/min$	MAP < 70	DOPAM. ≤ 5 OR DOBUTAM.	DOPAM. > 5 OR EPINEPH. ≤ 0.1 OR NOREPIN. ≤ 0.1	DOPAM. > 15 OR EPINEPH. > 0.1 OR NOREPIN. > 0.1
CENTRAL NERVOUS SYSTEM	GCS		13-14	10-12	6-9	<6
RENAL	CREATININE	MG/DL OR ML/D	1.2-1.9	2.0-3.4	3.5-4.9 OR URIN < 500	>5.9

Table 8: **Overview of the SOFA score subscores** (Wendland et al., 2024) Each column corresponds to a subscore (0–4) reflecting the severity of dysfunction in the central nervous, cardiovascular, respiratory, coagulation, hepatic, and renal systems. Subscores of 0 indicate normal organ function as defined by the respective criteria.

We consider patients treated with the antibiotics vancomycin, ceftriaxone, and piperacillin-tazobactam. To assess potential treatment side effects we additionally predict laboratory indicators of organ toxicity, including creatinine (kidney injury) as well as total bilirubin and alanine aminotransferase (ALT) (liver injury).

Data preprocessing follows the OptAB pipeline (Wendland et al., 2024). The cohort is restricted to ICU patients, and antibiotic treatments are encoded as a vector of three binary variables indicating administration of vancomycin, ceftriaxone, and piperacillin-tazobactam.

Time-dependent covariates were selected from all available dynamic variables by excluding variables recorded in fewer than 50% of patients and retaining those with an absolute Spearman correlation of at least 0.3 with the SOFA score during the first 12 hours. The resulting covariates are listed in Table 9 .

To capture information about measurement patterns, cumulative missingness indicators were added as additional time-dependent covariates. Static covariates include sex, age, height, and weight at admission.

The final dataset contains 3,556 patients. We use 80% of the data for training and 20% for testing, with the training set further split into 80% training and 20% validation data.

AmsterdamUMCdb sepsis cohort: For external validation, we evaluated the performance of ObsNODE, trained on MIMIC-IV, in the independent AmsterdamUMCdb cohort (Thoral et al., 2021). AmsterdamUMCdb comprises Electronic Health Records (EHR) data from 23,106 ICU admissions at Amsterdam University Medical Center between 2003 and 2016. The distribution of administered antibiotics differed substantially between AmsterdamUMCdb and MIMIC-IV, potentially reflecting differences in patient populations, pathogen prevalence, or local treatment practices. Among 6,285 patients with sepsis in AmsterdamUMCdb, 3,374 received Ceftriaxone, 3,058 Cefotaxime, and 1,718 erythromycin. For external validation, we restricted the cohort to patients treated with antibiotics Vancomycin, Ceftriaxone, or Piperacillin/Tazobactam. This resulted in an external test set of 399

VARIABLE	UNIT
SOFA SCORE	-
ALANINE AMINOTRANSFERASE	IU/L
ANION GAP	MEQ/L
BICARBONATE	MEQ/L
TOTAL BILIRUBIN	MG/DL
BLOOD UREA NITROGEN (BUN)	MG/DL
CREATININE	MG/DL
DIASTOLIC BLOOD PRESSURE	MMHG
PLATELET COUNT	K/μL
RED CELL DISTRIBUTION WIDTH (RDW)	%
SYSTOLIC BLOOD PRESSURE	MMHG

Table 9: Time-dependent covariates from MIMIC-IV-sepsis dataset

patients, comprising 364 treated with Ceftriaxone and 35 with vancomycin; no patients treated with Piperacillin/Tazobactam.

The two cohorts differed substantially in both SOFA score distributions and the availability of laboratory measurements. The mean SOFA score among patients with sepsis was higher in AmsterdamUMCdb than in MIMIC-IV (6.8 to 4.43). Laboratory measurements contributing to the SOFA score were also more frequently available in AmsterdamUMCdb. In the AmsterdamUMCdb test set, 23.3% of patients had no recorded total bilirubin measurement and 16.8% had no recorded alanine aminotransferase (ALT) measurement, compared with 56.8% and 56.1%, respectively, in MIMIC-IV. These differences indicate substantial heterogeneity in laboratory testing and documentation practices, which may contribute to the observed distribution shift and affect model robustness and generalizability across cohorts.

The AmsterdamUMCdb data were preprocessed following the OptAB pipeline (Wendland et al., 2024), using the same preprocessing procedures as applied to the MIMIC-IV dataset to ensure consistency between the training and external validation cohorts.

Data availability: The public MIMIC-IV dataset (Johnson et al., 2023; Goldberger et al., 2000) is available under `https://physionet.org/content/mimiciv/2.2/` and is shared by the data owners after reasonable request.

AmsterdamUMCdb dataset (Thoral et al., 2021) is available under `https://amsterdammedicaldatascience.nl/amsterdamumcdb/`. The data is shared by the data owners after reasonable request including the supervision of a practising intensivist.

A.5 Hyperparameter Tuning

A.5.1 Hyperparameter Tuning of ObsNODE

For hyperparameter optimization of the ObsNODE model, we use Optuna with the Tree-structured Parzen Estimator (TPE). Table 10 summarizes the search space (a) and the selected configurations for three datasets: the synthetic cancer dataset, the semi-synthetic MIMIC-IV dataset, and the MIMIC-IV sepsis dataset (b). While the parameters of the NODE component can, in principle, be tuned separately for each neural network in the observation normal form (7), we keep them identical across all components in our experiments.

SEARCH SPACE/DATA SET	SYNTHETIC CANCER	(SEMI-SYNTHETIC) MIMIC-IV
LATENT SPACE DIMENSION	$[d_y, .., 3\,d_y]$	$[d_y, .., 3\,d_y]$
BATCH SIZE	{32,64,...,256}	[100,250,500,750,1000]
LEARNING RATE	[0.001, 0.0001, 0.00001]	[0.001, 0.0001, 0.00001]
HIDDEN DIM OBSERVER	[32,64,128,256]	[32,64,128,256]
HIDDEN DIMENSION NODE	[32,64,128,256]	[32,64,128,256]
NUMBER OF LAYERS NODE	{1,2,...,10}	{1,2,...,10}
ACTIVATION FUNCTION NODE	[LEAKYRELU, TANH, SIGMOID]	[LEAKYRELU, TANH, SIGMOID]

(a) Search space for ObsNODE hyperparameters.

HYPERPARAMETER/DATA SET	SYNTHETIC CANCER	SEMI-SYNTHETIC MIMIC-IV	MIMIC-IV
LATENT SPACE DIMENSION = $d_y \times m$	2×2	1×2	11×3
BATCH SIZE	64	100	500
LEARNING RATE	0.001	0.0001	0.0001
HIDDEN DIMENSION OBSERVER	128	128	256
HIDDEN DIMENSION NODE	128	128	256
NUMBER OF LAYERS NODE	5	6	7
ACTIVATION FUNCTION NODE	LEAKYRELU	LEAKYRELU	LEAKYRELU

(b) Optimized hyperparameter values for ObsNODE.

Table 10: **Hyperparameter Tuning of the ObsNODE Model.** Table (a) illustrates the hyperparameter search space, while Table (b) reports the optimal hyperparameter configurations for experiments conducted on the synthetic cancer dataset, the semi-synthetic MIMIC-IV sepsis dataset, and the MIMIC-IV sepsis dataset.

A.5.2 Hyperparameter Tuning of OptAB and DoseAI

For hyperparameter tuning of the OptAB and DoseAI models, we use Optuna with the Tree-structured Parzen Estimator (TPE). Table 11 summarizes the search space (a) and the resulting configurations (b) for both models. DoseAI can handle continuous treatment doses and is applied to the synthetic cancer dataset, while OptAB is restricted to categorial treatments and used for the semi-synthetic dataset based on MIMIC-IV and the real world sepsis dataset from MIMIC-IV. The decoder optimization is initialized with the final hyperparameter values obtained for the encoder.

Search Space/Data set	synthetic cancer	(semi-synthetic) MIMIC-IV
Dimension of the latent state	{1, 2, ..., 30}	{1, 2, ..., 30}
batch size	{16, 32, 64, 125, 250, 500, 1000}	{100, 200, 500, 1000, 2000}
Max. number of NCDE units	{1, 2, ..., 1000}	{1, 2, ..., 1000}
Learning rate	[0.0001, 0.01]	[0.0001, 0.01]
Activation function of F	{leakyrelu, tanh, sigmoid, identity}	{leakyrelu, tanh, sigmoid, identity}
Number of NCDE layers	{1, 2, ..., 20}	{1, 2, ..., 20}
Activation function of h_α	{leakyrelu, tanh, sigmoid, identity}	{leakyrelu, tanh, sigmoid, identity}
Number of layers of h	{1, 2, ..., 1000}	{1, 2, ..., 1000}
Max. number of units of h	{1, 2, ..., 6}	{1, 2, ..., 6}

(a) Search space for OptAB and DoseAI hyperparameters.

Hyperparameter/Data set	synthetic cancer	semi-synthetic MIMIC-IV	MIMIC-IV
encoder			
Dimension of the latent state	16	7	17
batch size	250	1000	500
Max. number of NCDE units	578	174	33
Learning rate	0.0042	0.0016	0.0051
Activation function of F	leakyrelu	leakyrelu	tanh
Number of NCDE layers	2	14	15
Activation function of h_α	tanh	leakyrelu	tanh
Number of layers of h	128	275	128
Max. number of units of h	4	4	1
decoder			
Dimension of the latent state	22	25	17
batch size	125	1000	1000
Max. number of NCDE units	802	825	33
Learning rate	0.0016	0.0007	0.0051
Activation function of F	leakyrelu	identity	tanh
Number of NCDE layers	13	9	15
Activation function of h_α	leakyrelu	leakyrelu	tanh
Number of layers of h	798	403	128
Max. number of units of h	1	1	1

(b) Optimized hyperparameter values for OptAB and DoseAI Models.

Table 11: **Hyperparameter Tuning of the OptAB and DoseAI Models.** Table (a) presents the hyperparameter search space, while Table (b) reports the optimal hyperparameter configurations. DoseAI is used for training on the synthetic cancer dataset, whereas OptAB is applied to the semi-synthetic data set and the real world sepsis dataset from MIMIC-IV.

A.5.3 Hyperparameter Tuning of IGC-Net

For hyperparameter tuning of the IGC-Net model, we employ the Hydra implementation of the random grid search strategy for computational effciency reasons. Table 10 presents the hyperparameter search space (a) and the selected hyperparameter configuration (b) of the IGC-Net model.

SEARCH SPACE/ DATA SET	SYNTHETIC CANCER	SEMI-SYNTHETIC MIMIC-IV	MIMIC-IV
TRANSFORMER BLOCKS	[1,2]	1	1
LEARNING RATE (η)	[0.01, 0.001, 0.0001]	[0.001, 0.0001, 0.0001]	[0.001, 0.0001]
MINIBATCH SIZE	[64, 128, 256]	[32,64,128, 256,512]	[32,64]
ATTENTION HEADS	1	[1,2,3]	[2,3]
TRANSFORMER UNITS	$[1d_g, 2d_g, 3d_g, 4d_g]$	$[1d_g, 2d_g, 3d_g, 4d_g]$	$[0.5d_g, 1d_g]$
HIDDEN REPRESENTATION SIZE	$[0.5d_g, 1d_g, 2d_g, 3d_g, 4d_g]$	$[1d_g, 2d_g, 3d_g, 4d_g]$	$[0.5d_g, 1d_g, 2d_g]$
FEED-FORWARD HIDDEN UNITS	$[0.5d_z, 1d_z, 2d_z, 3d_z, 4d_z]$	$[1d_z, 2d_z, 3d_z, 4d_z]$	$[0.5d_z, 1d_z, 2d_z]$
SEQUENTIAL DROPOUT RATE	[0.1, 0.2]	[0.1, 0.2]	[0.1, 0.2]
MAX POSITIONAL ENCODING	15	30	15
NUMBER OF EPOCHS	50	100	100

(a) Search space for IGC-Net hyperparameters.

HYPERPARAMETER/DATA SET	SYNTHETIC CANCER	SEMI-SYNTHETIC MIMIC-IV	MIMIC-IV
TRANSFORMER BLOCKS	2	2	1
LEARNING RATE	0.01	0.01	0.001
MINIBATCH SIZE	64	64	64
ATTENTION HEADS	0.1	1	2
TRANSFORMER UNITS	32	24	30
HIDDEN REPRESENTATION SIZE	64	72	30
FEED-FORWARD HIDDEN UNITS	32	18	30
SEQUENTIAL DROPOUT RATE	0.1	0.1	0.1
MAX POSITIONAL ENCODING	15	30	15
NUMBER OF EPOCHS	50	100	100

(b) Optimized hyperparameter values for IGC-Net.

Table 12: **Hyperparameter Tuning of the IGC-Net Model.** Table (a) presents the hyperparameter search space, while Table (b) reports the optimal hyperparameter configurations of the IGC-Net model for experiments conducted on the synthetic cancer dataset, the semi-synthetic MIMIC-IV dataset, and the real world MIMIC-IV sepsis dataset. The overall input dimension is given by $d_g = d_y + d_x + d_a$, and d_z denotes the hidden representation size of the IGC-Net model.

A.5.4 Hyperparameter Tuning of SCIP-Net

For hyperparameter tuning of the SCIP-Net model for the synthetic cancer data and semi-synthetic data, we employed the Hydra–Optuna implementation of the Tree-structured Parzen Estimator (TPE). For the real sepsis data, we tuned the SCIP-Net model using a random grid search. Hyperparameter optimization with Optuna using the Tree-structured Parzen Estimator (TPE) did not yield improved or optimal parameter configurations.

Table 13 summarizes the hyperparameter search space (a) and the selected configurations (b) for the synthetic cancer dataset, the semi-synthetic MIMIC-IV sepsis dataset, and the MIMIC-IV sepsis dataset.

Search Space Model	Weight network, Treatment network, Encoder	Decoder
Neural CDE hidden layers	[1,2]	[1,2]
Learning rate	[0.01,0.001, 0.0001, 0.0001]	[0.01,0.001, 0.0001, 0.0001]
Minibatch size	[64,128,256]	[256,512,1024]
Neural CDE hidden units	[0.5,1,2,3,4]d_g	[1,2,3,4,8,16]d_g
Neural CDE dropout rate	[0.1,0.2]	[0.1,0.2]
Max. gradient norm	[0.5,1,2]	[0.5,1.0,2.0,4.0]
Epochs	50	50

(a) Search space for SCIP-Net hyperparameters.

Parameter / Model	Weight network	Treatment network	Encoder	Decoder
Synthetic cancer				
Neural CDE hidden layers	2	1	1	2
Learning rate	0.0001	0.01	0.01	0.001
Minibatch size	128	64	64	256
Neural CDE hidden units	8	12	8	32
Neural CDE dropout rate	0.1	0.1	0.1	0.1
Max. gradient norm	0.5	0.5	0.5	1
Semi-synthetic MIMIC-IV				
Neural CDE hidden layers	1	2	2	2
Learning rate	0.01	0.01	0.01	0.0001
Minibatch size	64	64	256	256
Neural CDE hidden units	9	12	6	12
Neural CDE dropout rate	0.1	0.1	0.2	0.2
Max. gradient norm	0.5	0.5	2	1
MIMIC-IV				
Neural CDE hidden layers	2	2	1	2
Learning rate	0.01	0.01	0.01	0.01
Minibatch size	256	256	256	512
Neural CDE hidden units	42	42	42	84
Neural CDE dropout rate	0.2	0.1	0.1	0.1
Max. gradient norm	0.5	2.0	2.0	0.5

(b) Optimized hyperparameter values for SCIP-Net.

Table 13: **Hyperparameter Tuning of the SCIP-Net Model.** Table (a) presents the hyperparameter search space. Table (b) reports the optimal configurations for the synthetic cancer dataset and the semi-synthetic MIMIC-IV sepsis dataset and the real world MIMIC-IV sepsis dataset.

A.5.5 Hyperparameter Tuning of NODE-Model

As a baseline model, we implemented a Neural Ordinary Differential Equation (NODE) model using the same encoder-LSTM architecture as employed for the ObsNODE model. In contrast to the ObsNODE formulation, the latent dynamics are modeled using a standard latent Neural ODE without incorporating an observability normal form. The latent dimension was chosen to be identical to that of the ObsNODE model to ensure a fair comparison between both approaches.

For the hyperparameter optimization of the NODE model, we employed Optuna with the Tree-structured Parzen Estimator (TPE) sampler. The considered search space and the resulting optimal hyperparameter configuration are summarized in Table 14.

Search Space/ Data set	synthetic cancer	semi-synthetic MIMIC-IV
Batch size	{32,64,...,256}	[100,250,500,750,1000]
Learning rate	[0.001, 0.0001, 0.00001]	[0.001, 0.0001, 0.00001]
Hidden dimension Observer	[32,64,128,256]	[32,64,128,256]
Hidden dimension NODE	[32,64,128,256]	[32,64,128,256]
Number of layers NODE	{1,2,...,10}	{1,2,...,10}
Activation function NODE	[leakyrelu, tanh, sigmoid]	[leakyrelu, tanh, sigmoid]

(a) Search space for NODE hyperparameters.

Hyperparameter/ Data set	synthetic cancer	semi-synthetic MIMIC-IV
Batch size	96	100
Learning rate	0.001	0.001
Hidden dimension Observer	256	64
Hidden dimension NODE	64	64
Number of layers NODE	8	3
Activation function NODE	sigmoid	leakyrelu

(b) Optimized hyperparameter values for NODE.

Table 14: **Hyperparameter tuning of the NODE model.** Hyperparameter search space (a) and optimal configuration (b) for the NODE model on the synthetic cancer and semi-synthetic MIMICIV-sepsis dataset.

A.6 Computational efficiency

For training and hyperparameter optimization of ObsNODE we utilized two Nvidia A100 GPUs, each with 40 GB of memory. The same setting was used for the baseline models DoseAI, OptAB and SCIP-Net. For IGC-Net, we used a NVIDIA H100 GPU with 96 GB of memory.

Training of ObsNODE for the optimal hyperparameter values on a single Nvidia A100 GPU took 140 min for the synthetic cancer data, 49 min for the semi-synthetic data and 65 min for the real world sepsis data.

This is similar to the training times required for the baseline models IGC-Net and SCIP-Net. Training OptAB and DoseAI required between 40 and 74 hours, depending on the data set.

A.7 Declaration of LLM usage

The core method development in this research does not involve LLMs as any important original, or non-standard components.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the main contributions of the paper: (i) a theoretical result establishing that observability of the latent state is a necessary condition for identifying dynamic treatment effects in continuous-time state-space models with hidden confounding, together with the derivation of a corresponding causal adjustment formula, (ii) the proposed ObsNODE framework, which enforces observability via an observable normal form, and (iii) empirical evaluation on synthetic, semi-synthetic, and real-world datasets demonstrating strong causal forecasting performance compared to existing methods. The scope and limitations are also appropriately reflected: the approach relies on assumptions about the data-generating process, and results are validated on controlled and real-world clinical datasets without claiming universal applicability. Overall, the claims in the abstract and introduction are consistent with the theoretical and empirical results presented in the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses the main limitations in the final section. In particular, we assume that the data-generating process can be described by a continuous-time structural causal model with latent state dynamics governed by a process which can be described by an ODE. This implies that patient trajectories are governed by an underlying latent disease state. The framework allows for hidden time-dependent confounders acting on both treatment and outcome, provided that their effects are not directly acting on the latent state, see Figure 1. Observability is not assumed ad hoc but enforced by the ObsNODE architecture through the observable normal form. Finally, the empirical evaluation is limited to synthetic, semi-synthetic, and real-world clinical datasets, and the proposed adjustment formula has not yet been fully exploited for uncertainty quantification, which remains an important direction for future methodological work.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.

- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The paper provides a complete set of assumptions for all theoretical results, including the latent state-space formulation and the structural causal data-generating process. All assumptions are clearly stated and linked to the identifiability result and the derivation of the causal adjustment formula. The main theoretical results are formally stated. The paper includes an intuitive proof sketch in the main text based on a discrete-time approximation, while complete and rigorous proofs are provided in the appendix. The proofs rely on standard concepts from control theory and causal inference, which are appropriately referenced. Overall, the assumptions and proofs are clearly specified and the theoretical development is complete.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides sufficient detail to reproduce the main experimental results. The ObsNODE architecture is fully specified, including the continuous-time state-space formulation, observable normal form, and treatment modeling. The training procedure, loss functions, and optimization setup are described, along with relevant hyperparameters. The datasets used for evaluation (synthetic cancer data, semi-synthetic MIMIC-IV data, and real-world sepsis data) are clearly specified, and the construction of the semi-synthetic datasets is described. Evaluation metrics, prediction horizons, and experimental protocols (including repeated runs and reporting of mean and standard deviation) are explicitly stated. Together, these details enable independent reproduction of the reported results and verification of the main empirical claims.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The paper provides open access to a GitHub repository containing the implementation of ObsNODE, along with data simulation and preprocessing scripts and detailed instructions to reproduce the main experimental results. The semi-synthetic data generation pipeline is fully included, enabling reconstruction of these experiments. Due to data usage restrictions, the MIMIC-IV dataset cannot be redistributed; however, it is publicly available via an official access procedure, and the corresponding link and instructions are provided in the appendix. The repository includes environment specifications and example commands for running the experiments. Overall, the provided code, scripts, and documentation offer a clear and practical pathway to reproduce the reported results.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).

- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies all relevant experimental details required to understand and reproduce the results. The main text describes the datasets (synthetic, semi-synthetic MIMIC-IV, and real-world MIMIC-IV), evaluation protocol, and performance metrics. Data splits, preprocessing (including normalization), and training procedures are clearly defined. Model architectures and training strategies, including the Neural ODE formulation, are described in detail. Hyperparameter search spaces and selected configurations for all models are provided in the appendix, along with the optimization procedure and training settings (e.g., optimizer, batch sizes). Additional details on data generation, implementation, and computational setup are also included in the appendix, ensuring full transparency.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Experimental results are reported with mean and standard deviation over five independent runs with different train/validation/test splits. RMSE values in tables are given as mean $\pm$ standard deviation, and heatmaps are averaged over runs. The reported variability therefore captures both data split and stochastic training effects. The definition of the error measure (RMSE) and aggregation procedure are clearly described in the text. While no formal hypothesis tests are performed, the consistent reporting of standard deviations across all main experiments provides a clear assessment of statistical variability and robustness.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper reports the computational resources used for all experiments. Training and hyperparameter optimization were conducted on NVIDIA A100 GPUs (40 GB memory), with IGC-Net trained on an NVIDIA H100 GPU (96 GB memory). Training times for ObsNODE are provided for each dataset (synthetic, semi-synthetic, and real-world), and comparable runtimes for baseline models are discussed. This information allows estimation of the computational cost per run and highlights differences in efficiency across methods. While the total compute budget for all exploratory experiments is not explicitly quantified, the reported hardware specifications and per-run training times provide sufficient detail to reproduce the main experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research complies with the NeurIPS Code of Ethics. All experiments are conducted on synthetic, semi-synthetic, and publicly available clinical data (MIMIC-IV) with appropriate access controls. No identifiable personal data are disclosed, and all analyses follow established data usage agreements. The work is presented transparently, including assumptions, limitations, and potential sources of bias. Code and experimental details are provided to support reproducibility, and no harmful or deceptive applications are pursued.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [No]

Justification: The paper does not include a dedicated discussion of broader societal impacts. The proposed method is motivated by clinical decision support and aims to improve individualized treatment effect estimation, which may have positive implications for patient care. At the same time, as with all models trained on observational healthcare data, there is a risk of bias due to data quality and confounding, which could affect downstream decisions if used without proper validation. While the causal formulation may help mitigate spurious associations under appropriate assumptions, a detailed analysis of societal impacts and risks is left for future work.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not introduce assets with a high risk of misuse. The proposed method is a modeling framework for causal forecasting and does not involve releasing generative models, scraped datasets, or other artifacts with dual-use concerns. Experiments are conducted on synthetic, semi-synthetic, and controlled-access clinical data (MIMIC-IV), and the released code does not pose inherent safety risks. Therefore, no specific safeguards are required.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [No]

    Justification: The paper properly credits all datasets, models, and prior work used in the experiments, including the MIMIC-IV dataset and baseline methods, and provides references and URLs where appropriate. However, the specific licenses and terms of use (e.g., data access agreements for MIMIC-IV) are not explicitly stated in the paper. All data usage complies with the respective access requirements, and future versions will include explicit license information for all external assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper introduces new assets, including the implementation of ObsNODE and scripts for data generation, preprocessing, and experiments, which are made available via a public repository. The appendix provides detailed descriptions of the data generation processes, model architecture, and training procedures. While the code repository contains the necessary components to reproduce the experiments, documentation is primarily provided through the paper and appendix rather than a fully structured artifact description (e.g., model cards or detailed usage guidelines), which could be extended in future work.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

Justification: The paper does not involve crowdsourcing or research with human subjects. All experiments are conducted using synthetic, semi-synthetic, and publicly available clinical datasets (MIMIC-IV), and no human participants were recruited or directly involved in the study.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve research with human subjects. All experiments are conducted using synthetic, semi-synthetic, and publicly available de-identified clinical datasets (MIMIC-IV). No interaction with human participants was performed, and no individual-level identifiable data were used.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this research does not involve large language models as part of the methodology. LLMs are not used in model design, training, or evaluation. Any use of LLMs was limited to writing or editing support and does not affect the scientific contributions or results.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.