
DoseSurv: Predicting Personalized Survival Outcomes under Continuous-Valued Treatments

Moritz Gögl¹ Yu Liu¹ Christopher Yau^{1,2} Peter Watkinson¹ Tingting Zhu¹

¹University of Oxford

²Health Data Research UK

moritz.gogl@keble.ox.ac.uk

Abstract

Estimating heterogeneous treatment effects (HTEs) of continuous-valued interventions on survival, that is, time-to-event (TTE) outcomes, is crucial in various fields, notably in clinical decision-making and in driving the advancement of next-generation clinical trials. However, while HTE estimation for continuous-valued (i.e., dosage-dependent) interventions and for TTE outcomes have been separately explored, their combined application remains largely overlooked in the machine learning literature. We propose DoseSurv, a varying-coefficient network designed to estimate HTEs for different dosage-dependent and non-dosage treatment options from TTE data. DoseSurv uses radial basis functions to model continuity in dose-response relationships and learns balanced representations to address covariate shifts arising in HTE estimation from observational TTE data. We present experiments across various treatment scenarios on both simulated and real-world data, demonstrating DoseSurv’s superior performance over existing baseline models.

1 Introduction

Estimating the causal effects of interventions at the individual level is essential for informed decision making in fields such as personalized medicine [48], public health [39], social and economic policy [62], and marketing and retail [44]. Within the machine learning (ML) literature, predominant attention has been directed toward estimating heterogeneous treatment effects (HTEs) on continuous and binary outcomes [1, 2, 62, 70, 63, 60, 9, 16]. By contrast, time-to-event (TTE) outcomes, which capture the time until an event of interest, have received comparatively less attention. More specifically, the existing ML literature on HTE estimation with TTE outcomes has focused on quantifying causal effects of binary (non-dosage) interventions with “treatment” or “no treatment” (control) being the only options [14, 58]. However, in practice, treatment settings are typically more complex, comprising several, possibly continuous-valued treatment options, characterized by their dosage or frequency of administration.

TTE outcomes have immense practical relevance, for example, in clinical studies that analyze survival time or time to discharge. Equally, studying the causal effects of continuous-valued treatments is an important problem in healthcare, with relevant areas including chemotherapy [18], insulin regimens [36], and cardiovascular treatments [21]. Underlying dose-response relationships can be complex and non-monotonic, with multiple stimulatory and inhibitory phases. Although average dose-response curves across a population can be characterized in randomized controlled trials (RCTs) and may appear simple, dose-response relationships at the patient level are often more intricate. Individual patient characteristics can complicate these relationships, leading to patient-specific ideal dosages and necessitating continuous modeling of treatments at the individual level.

While neural network (NN) models for HTE estimation have been developed separately for continuous-valued interventions [60, 9, 51] and for TTE outcomes [12, 14, 58], their joint application remains

largely unexplored in the current ML literature. Therefore, we propose *DoseSurv*, a deep learning survival model that estimates the conditional hazard function—the probability of experiencing the event at a given time—under different treatments and dosages. Unlike many (semi-)parametric survival models—such as the Cox proportional hazards model [13] and related treatment-effect models [58]—our model does not require covariate effects on the hazard to be constant over time.

Instead, we adopt NN-based discrete-time survival models that estimate conditional hazards through a series of binary classification tasks [10, 24], following their adaptation for HTE estimation with binary treatments by Curth et al. [14]. From the estimated conditional hazards, we can derive more practically relevant measures, such as treatment-specific survival functions and the (restricted) mean survival time, which in turn allow us to define treatment effects and dose-response relationships. The core component of *DoseSurv* is the radial basis function (RBF) hazard estimator, which models the survival hazard as a treatment-specific continuous function of dosage via Gaussian RBFs, leveraging the varying-coefficient network framework introduced by Nie et al. [51]. Through shared representations, *DoseSurv* also draws on information across different treatment options. Furthermore, it incorporates balanced representation learning via integral probability metric (IPM) regularization to mitigate various covariate shifts encountered in TTE and observational datasets [62, 14]. Here, event occurrence (survival), censoring (e.g., loss to follow-up), as well as treatment and dosage assignment, typically depend on individual patient characteristics leading to different covariate shifts [14]. We therefore evaluate *DoseSurv* on real and synthetic datasets spanning these covariate shifts and different treatment scenarios to assess its ability to mitigate the resulting biases.

Contributions Our contributions are fivefold: (i) We address the problem of HTE estimation for single- and multi-arm treatment regimens with mutually exclusive continuous-valued (dosage-dependent) and non-dosage treatment options, as well as TTE outcomes—a setting that has been largely overlooked despite its high practical relevance. (ii) We propose *DoseSurv*, a novel varying-coefficient network based on layer-specific RBFs and designed for HTE estimation in settings such as the one described above. (iii) We adopt a computationally efficient IPM regularization technique to mitigate various biases arising in HTE estimation from observational TTE data and continuous-valued treatments. (iv) Moreover, we create synthetic TTE datasets for different continuous-valued treatment types, and (v) we propose an extension of the classical *Twins* dataset, as a real-world benchmark for continuous-valued treatment settings with TTE data. Our results show that *DoseSurv* outperforms existing baselines under different treatment scenarios and covariate shifts.

2 Related Work

ML methods for estimating HTEs from binary interventions are well-studied and include model-agnostic meta-learners (e.g., S- and T-learners) [40, 47, 72] and model-specific approaches like multi-task architectures that leverage shared information across treatment groups while also allowing treatment-specific modeling [62, 16, 2]. Implementations typically span tree-based models [69, 6] and NNs [34, 62, 63, 16], and most works consider binary interventions only.

For continuous-valued interventions, classical approaches use the generalized propensity score [30], while recent NN-based architectures directly model dose-response relationships. Some methods discretize dosage into bins with treatment-specific heads [60], offering regional flexibility at the expense of strict continuity; others adopt varying-coefficient formulations in which parameters change smoothly with dosage [51, 61], preserving continuity. Adversarial approaches have also been explored for continuous-valued interventions [9]; in TTE settings, however, censoring and temporal dependence complicate defining faithful discrimination targets and stable training objectives.

In survival analysis, traditional models such as the Cox proportional hazards model [13] have been extended using NNs [20, 37]. Non-parametric alternatives include random survival forests [32] and discrete-time NN approaches [10, 45, 24], alongside survival clustering methods that group samples with similar risk profiles [33]. Integrating HTE estimation directly into survival models remains comparatively underexplored: most existing work addresses binary treatments using tree-based [29] or NN approaches [14, 58], with no support for continuous-valued treatments.

We provide an extended review of methods for survival analysis and HTE estimation—including those targeting continuous-valued interventions—in Appendix D, and summarize the applicability of related neural approaches across settings in Table D.1.

3 Problem Statement

3.1 Definitions

We aim to analyze an observational TTE dataset $\mathcal{D}$ comprising N individuals, each of whom is treated with one of M different continuous-valued treatments. Throughout, we largely follow notation and discrete-time survival formulations similar to those in [14]. The dataset can then be represented as $\mathcal{D} = \{\mathbf{x}_i, \tilde{\tau}_i, \delta_i, (a_i, q_i)\}_{i=1}^N$, where each patient i is a realization of random variables $(\mathbf{X}, \tilde{T}, \Delta, (A, Q)) \sim \mathbb{P}$. In this context, $\mathbf{X} \in \mathcal{X}$ refers to a vector of random variables associated with the patient covariates. Furthermore, we use $T \in \mathcal{T}$ and $C \in \mathcal{T}$ to represent the random variables for the event time (e.g., the time until death or onset of disease) and the censoring time (i.e., the time until loss of follow-up). The observed patient outcomes are given by $\Delta = \mathbb{1}(T \leq C)$, indicating whether either the event or censoring was observed, and $\tilde{T} = \min(T, C)$, representing the time of that observation. We denote the random variable for the treatment assignment as $A \in \mathcal{A} = \{0, \dots, M-1\}$. Each of these treatments $A \in \mathcal{A}$ is associated with a continuous random variable $Q \in \mathcal{Q} = [0, 1] \subset \mathbb{R}$, representing the (normalized) dosage for the respective treatment.¹ We treat time as discrete, with a finite time horizon t_m , such that $\mathcal{T} = \{1, \dots, t_m\}$. Based on the notation above, the probability of the event occurring at time τ for a given patient with covariates $\mathbf{x}$ and intervention² (a, q) who has not experienced an event prior to τ is defined as [14]

$$\lambda(\tau|\mathbf{x}, a, q) = \mathbb{P}(\tilde{T} = \tau, \Delta = 1 | \mathbf{X} = \mathbf{x}, A = a, Q = q, \tilde{T} \geq \tau), \quad (1)$$

which is also known as *conditional hazard*. From this, we can compute the survival function $S(\tau|\mathbf{x}, a, q) = \prod_{t \leq \tau} (1 - \lambda(t|\mathbf{x}, a, q))$, representing the probability that the event will not occur up until time τ . To infer the conditional hazards (and thus the survival function), we follow a typical approach in discrete-time survival analysis by estimating $\lambda(\tau|\mathbf{x}, a, q)$ from covariates and time intervals using a non-parametric ML model. We transform the TTE data into *person-period* (long) format [66] and treat the estimation of the discrete conditional hazard function as a series of binary classification problems associated with different time intervals. More specifically, we can fit an NN model to the binary event indicator at each discrete time interval, with $\mathbf{x}$, a , q , and τ as predictors.

3.2 Relationship to Potential Outcomes Framework

In HTE estimation, our aim is not primarily to estimate outcomes conditional on the intervention, as presented above. Instead, we want to estimate the *potential outcome* that will occur in the event that a particular action is performed. Following the Rubin causal model [31], we can adopt the survival function as a potential outcome of interest under a potential intervention (a, q) [14]

$$S_{a,q}(\tau|\mathbf{x}) = \mathbb{P}(T > \tau | \mathbf{X} = \mathbf{x}, \text{do}(A = a, Q = q, C \geq \tau)), \quad (2)$$

where the do-operator [53] shows that a specific action is being taken. Here, we have to treat censoring as an action we can intervene on by setting $C \geq \tau$, so that potential outcomes reflect uncensored event times and causal effects are identifiable [64].

The HTE between two potential interventions (a, q) and $(\tilde{a}, \tilde{q})$ may then, for example, be defined as the (time-dependent) difference in survival probabilities $\nu_{\text{surv}}(\tau|\mathbf{x}) = S_{a,q}(\tau|\mathbf{x}) - S_{\tilde{a},\tilde{q}}(\tau|\mathbf{x})$. Alternatively, we may compute the restricted mean survival time (RMST) until the time horizon t_m under an intervention (a, q) as a time-independent potential outcome of interest: $t_{a,q}^{\text{rmst}}(\mathbf{x}) = \sum_{t=0}^{t_m-1} S_{a,q}(t|\mathbf{x})$. The HTE can then be defined as $\nu_{\text{rmst}}(\mathbf{x}) = t_{a,q}^{\text{rmst}}(\mathbf{x}) - t_{\tilde{a},\tilde{q}}^{\text{rmst}}(\mathbf{x})$ [14]. Moreover, for a fixed treatment a , the function $q \mapsto t_{a,q}^{\text{rmst}}(\mathbf{x})$ defines a patient-specific dose-response curve that summarizes expected survival as a function of dosage.

To identify and quantify HTEs and dose-responses from TTE data, we make the following *assumptions*, which are standard in the literature on (dosage-dependent) treatment effect estimation [9, 60, 14] (Assumptions 1–3) and TTE outcomes [19, 50, 14] (Assumptions 4–5), and extend them for continuous-valued treatments. We denote the potential event time under a potential intervention (a, q) as $T_{a,q}$.

¹For simplicity, we focus here on a set of continuous-valued treatments only, i.e., $\mathcal{A} = \mathcal{A}^{\text{cont}}$. However, DoseSurv can also accommodate the more general case with continuous-valued (dosage-dependent) and categorical (non-dosage) treatment options, i.e., $\mathcal{A} = \mathcal{A}^{\text{cont}} \cup \mathcal{A}^{\text{cat}}$; for the latter, q is a dummy with no effect.

²In the following, we refer to the tuple (a, q) as “intervention” to distinguish it from a “treatment” a with associated “dosage” q .

Assumption 1 (Unconfoundedness) For all $a \in \mathcal{A}$ and $q \in \mathcal{Q}$, the intervention and the potential event time are conditionally independent given covariates: $(A, Q) \perp T_{a,q} \mid \mathbf{X}$.

Assumption 2 (Positivity of Interventions) For all $\mathbf{x} \in \mathcal{X}$ and all $a \in \mathcal{A}$:

- (2a) *Positivity of Treatment*: There exists $\epsilon > 0$ such that $\epsilon < \mathbb{P}(A = a \mid \mathbf{X} = \mathbf{x}) < 1 - \epsilon$.
- (2b) *Positivity of Dosage*: For any admissible dose $q \in \mathcal{Q}$ and any small interval $\mathcal{U} \subset \mathcal{Q}$ around q , $\mathbb{P}(Q \in \mathcal{U} \mid A = a, \mathbf{X} = \mathbf{x}) > 0$.

Assumption 3 (Consistency) The potential event time will actually be observed under the observed intervention without being affected by any other external factors, i.e., if $A = a$ and $Q = q$, then $T = T_{a,q}$.

Assumption 4 (Independent Censoring) For all $a \in \mathcal{A}$, $q \in \mathcal{Q}$, the potential event time and the censoring time are conditionally independent given covariates and intervention: $T_{a,q} \perp C \mid \mathbf{X}, (A, Q)$.

Assumption 5 (Positivity of Events and Censoring) For all $\mathbf{x} \in \mathcal{X}$, $a \in \mathcal{A}$, $q \in \mathcal{Q}$, and all $t \in \mathcal{T}$:

- (5a) *Positivity of Events*: $\mathbb{P}(T > t \mid \mathbf{X} = \mathbf{x}, A = a, Q = q) \geq \epsilon$ for some $\epsilon > 0$.
- (5b) *Positivity of Censoring*: $\mathbb{P}(C > t \mid \mathbf{X} = \mathbf{x}, A = a, Q = q) \geq \epsilon$ for some $\epsilon > 0$.

Under Assumptions 1, 3, and 4, we can identify the interventional hazard from the observational one, i.e., $\lambda(\tau \mid \mathbf{x}, a, q) = \lambda_{a,q}(\tau \mid \mathbf{x}) = \mathbb{P}(T = \tau \mid T \geq \tau, \mathbf{X} = \mathbf{x}, \text{do}(A = a, Q = q, C \geq \tau))$. The corresponding potential outcome survival function can then be computed as $S_{a,q}(\tau \mid \mathbf{x}) = \prod_{t \leq \tau} (1 - \lambda_{a,q}(t \mid \mathbf{x}))$. A brief mathematical derivation of this identification is given in Appendix F. Assumptions 2 and 5 further provide the overlap and non-degeneracy conditions for treatment, dosage, and at-risk sets required for non-parametric estimation [14]. We provide the causal diagram underlying the assumed data-generating process in Appendix G.

4 DoseSurv

4.1 Model Architecture

Based on the definition of the problem presented above, we propose an NN-based discrete-time survival model, called DoseSurv, which is depicted in Fig. 1, and is tailored for survival predictions in single- and multi-arm treatment regimens with dosage-dependent (and non-dosage) treatments. DoseSurv consists of a *representation network*, $\Phi : \mathcal{X} \rightarrow \mathcal{R}$, which is implemented as a fully connected NN and shared across all treatments a , dosages q , and times τ . It is followed by treatment-specific *hazard estimators*, implemented as individual networks, each performing multi-output binary classification $h^{(a)} : \mathcal{R} \times \mathcal{Q} \rightarrow [0, 1]^{t_m}$, which allow us to estimate the survival hazards at each time point τ for a specific continuous-valued treatment a

$$\hat{\lambda}(\tau \mid \mathbf{x}, a, q) = h_{\tau}^{(a)}(\Phi(\mathbf{x}), q), \quad (3)$$

where $h_{\tau}^{(a)}$ is the output of $h^{(a)}$ corresponding to time τ . The treatment-specific estimators enable the learning of features unique to each respective treatment option. To enable the learning of the dosage dependency in the hazard estimates, we take two steps: (1) we add q as an additional feature to the shared representation $\Phi(\mathbf{x})$. (2) more importantly, we use Gaussian RBFs to explicitly model the network parameters (weights and biases) as continuous functions of q , as described below.³

4.2 The RBF Hazard Estimator

For continuous-valued treatments a , we want to preserve the prominent role of q and prevent its influence from vanishing in a high-dimensional feature space. Therefore, we implement the treatment-specific hazard estimators as flexible varying-coefficient networks, where the network parameters θ are functions of dosage q , following prior work by Nie et al. [51]. We depart from that formulation by modeling the network parameters as linear combinations of treatment- and layer-specific RBFs with learnable centers and bandwidths. Concretely, let the hazard estimator

³For non-dosage treatments $a \in \mathcal{A}^{\text{cat}}$, $h^{(a)} : \mathcal{R} \rightarrow [0, 1]^{t_m}$ is implemented as a standard fully connected NN.

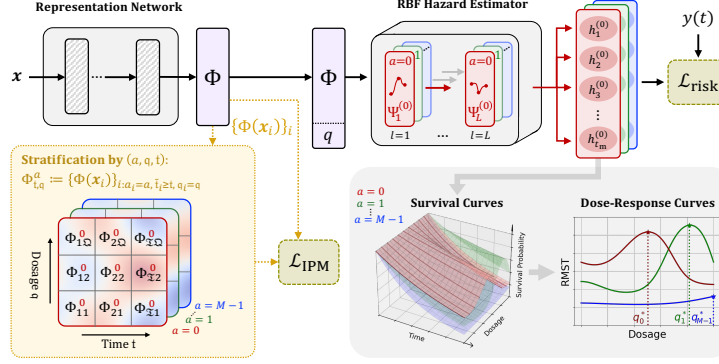


Figure 1: DoseSurv features a shared representation network that captures common features across treatments, dosages, and time intervals. To address covariate shifts arising from treatment-selection bias, dosage-selection bias, censoring, and event-induced shifts, we introduce an IPM regularization term based on a custom discretization of dosage ($q \rightarrow \hat{q}$) and time ($t \rightarrow \hat{t}$), enabling the learning of balanced representations Φ across treatments, dosages, and event times. For continuous treatments, $a \in \mathcal{A}^{\text{cont}}$, the model extends to the RBF hazard estimator, which consists of treatment-specific network heads. These heads use treatment- and layer-specific RBF bases $\{\psi_{l,k}^{(a)}\}_{k=1}^K$ to model parameters as continuous functions of q , ensuring continuity of discrete-time hazard estimates with respect to dosage. The hazard estimates allow for the computation of treatment- and dosage-specific survival curves, enabling dose-response analysis.

$h^{(a)}$ comprise J_l parameters (weights and biases) in each layer $l \in \{1, \dots, L\}$, which are given by $\theta_l^{(a)}(q) = (\theta_{1,l}^{(a)}(q), \theta_{2,l}^{(a)}(q), \dots, \theta_{J_l,l}^{(a)}(q))^\top$. Each of these parameters, indexed by j , is modeled as a continuous function of q

$$\theta_{j,l}^{(a)}(q) = \sum_{k=1}^K \beta_{j,l,k}^{(a)} \psi_{l,k}^{(a)}(q), \quad (4)$$

where $\{\psi_{l,k}^{(a)}\}_{k=1}^K$ are the K basis functions in layer l for treatment a , and $\beta_{j,l,k}^{(a)}$ are the learnable coefficients [51]. We can therefore write

$$\theta_l^{(a)}(q) = B_l^{(a)} \Psi_l^{(a)}(q) \quad (5)$$

with

$$B_l^{(a)} = \begin{bmatrix} \beta_{1,l,1}^{(a)} & \dots & \beta_{1,l,K}^{(a)} \\ \vdots & \ddots & \vdots \\ \beta_{J_l,l,1}^{(a)} & \dots & \beta_{J_l,l,K}^{(a)} \end{bmatrix} \in \mathbb{R}^{J_l \times K}, \quad \Psi_l^{(a)}(q) = (\psi_{l,1}^{(a)}(q), \psi_{l,2}^{(a)}(q), \dots, \psi_{l,K}^{(a)}(q)) \in \mathbb{R}^K.$$

We propose the use of Gaussian RBFs as basis functions

$$\psi_{l,k}^{(a)}(q) = w_{l,k}^{(a)} \exp \left[- (q - c_{l,k}^{(a)})^2 / (2(\sigma_{l,k}^{(a)})^2) \right] \quad (6)$$

with learnable centers $(c_{l,1}^{(a)}, c_{l,2}^{(a)}, \dots, c_{l,K}^{(a)})$, shape parameters $(\sigma_{l,1}^{(a)}, \sigma_{l,2}^{(a)}, \dots, \sigma_{l,K}^{(a)})$ and additional weights $(w_{l,1}^{(a)}, w_{l,2}^{(a)}, \dots, w_{l,K}^{(a)})$.⁴ This flexibility, together with the universal approximation property of Gaussian RBF expansions in q , implies that—given sufficiently many centers and suitable bandwidths—our q -dependent parameterization can represent any continuous dose-response function on compact intervals [52]. Moreover, Gaussian RBFs provide global support over the dosage range, potentially aiding in low-overlap regions. Furthermore, because complex survival likelihoods and their gradients interact nonlinearly with basis functions, our approach aims to boost optimization by learning not only the varying coefficients but also the properties of the basis functions themselves.

⁴While not explored here, coupling these amplitude weights across clinically similar treatments would provide a simple route to induce soft cross-treatment information sharing between hazard estimators.

4.3 Loss Function and Representation Balancing

DoseSurv uses both empirical risk minimization and representation balancing to learn hazard estimates in the presence of biases in treatment and dosage selection, as well as censoring and event occurrence.

Empirical Risk Loss We minimize the empirical risk through the following loss term [24, 41]

$$\mathcal{L}_{\text{risk}} = \frac{1}{N} \frac{1}{t_m} \sum_{t=1}^{t_m} \sum_{i: \tilde{\tau}_i \geq t} \text{BCE}(y_i(t), h_t^{(a_i)}(\Phi(\mathbf{x}_i), q_i)), \quad (7)$$

where $y_i(t) = 1\{\delta_i = 1 \wedge \tilde{\tau}_i = t\}$ denotes the observed event indicator for patient i in person-period (long) format. In this formulation, $\mathcal{L}_{\text{risk}}$ is proportional to the negative discrete-time hazard log-likelihood—that is, the average binary cross-entropy (BCE) computed over at-risk rows only. Minimizing it corresponds to maximum-likelihood estimation of per-interval hazards, and censoring is handled via the at-risk conditioning $i: \tilde{\tau}_i \geq t$.

Representation Balancing To mitigate the effects of confounding, as well as censoring and event-induced shifts, DoseSurv aims to learn a representation $\Phi(\mathbf{x})$ in which features are balanced across at-risk populations at different times under different treatments and dosages. For this purpose, we adapt an IPM regularization scheme for binary treatments [14], to minimize the Wasserstein distance [68] between feature distributions of individual time-, treatment-, and dosage-specific subpopulations and the overall baseline population. We discretize the continuous dosage q into Ω equally-sized bins $\{[0, 1/\Omega), [1/\Omega, 2/\Omega), \dots, [(\Omega-1)/\Omega, 1]\}$, indexed by $q \in \{1, 2, \dots, \Omega\}$, and denote by q_i the dosage-bin index for sample i . To prevent oversegmentation into overly small subpopulations—which can reduce computational efficiency, generalizability, and stability—we also discretize time t into $\mathfrak{T}$ intervals of lower temporal resolution, indexed by $t \in \{1, 2, \dots, \mathfrak{T}\}$. Let $\tilde{t}_i$ be the coarse time-bin index for sample i , obtained by mapping the observed time $\tilde{\tau}_i$ to one of these $\mathfrak{T}$ intervals. We then define the IPM regularization loss as the sum of Wasserstein distances between the representations of the total baseline population and that of each subpopulation

$$\mathcal{L}_{\text{IPM}} = \frac{1}{|\mathcal{A}|} \frac{1}{\Omega} \frac{1}{\mathfrak{T}} \sum_{a \in \mathcal{A}} \sum_{q=1}^{\Omega} \sum_{t=1}^{\mathfrak{T}} \text{Wass}(\{\Phi(\mathbf{x}_i)\}_i^N, \{\Phi(\mathbf{x}_i)\}_{i: a_i=a, \tilde{t}_i \geq t, q_i=q}). \quad (8)$$

Consequently, $\mathcal{L}_{\text{IPM}}$ penalizes, for each treatment a , time bin t , and dosage bin q , the discrepancy in Φ between the corresponding at-risk subpopulation and the overall baseline population.⁵ In principle, one could complement or replace our IPM regularization with observation reweighting tailored to dynamic survival settings with continuous treatments. Such an approach would require estimating generalized propensity scores for treatment and dosage given covariates, as well as conditional at-risk (survival) probabilities over time. In practice, these quantities are difficult to estimate reliably—especially with continuous dosages and time-varying risk sets—and often yield extreme or unstable weights under limited overlap, which can inflate variance and destabilize training. To avoid these issues, we focus on IPM-based representation balancing.

We calculate the total loss function as the sum of the empirical loss and the IPM regularization term

$$\mathcal{L} = \mathcal{L}_{\text{risk}} + \gamma \mathcal{L}_{\text{IPM}}. \quad (9)$$

Here, γ is a hyperparameter that controls the influence of the regularization term. More details on the implementation of DoseSurv are provided in Appendix E.

5 Experiments

5.1 Experimental Setup

Following previous work on HTE estimation with continuous-valued treatments [60, 9] and TTE outcomes [14, 58], we first evaluate the performance of our model on synthetic data, which provides access to counterfactual outcomes and ground-truth survival curves for individual patients, and is therefore essential for a thorough evaluation of causal effect estimation. In addition, we evaluate DoseSurv on real-world data, namely the *Twins* dataset.

⁵For non-dosage treatments $a \in \mathcal{A}^{\text{cat}}$, stratification of at-risk populations reduces to (a, t) ; we retain the same normalization to keep IPM loss contributions comparable across continuous-valued and non-dosage arms.

5.1.1 Synthetic Datasets

We consider three simulated treatment scenarios (S1–S3) and create synthetic TTE datasets with different underlying outcome generation processes for each of them. Patient covariates are correlated in each of the datasets. In S1, we evaluate model performance for a single continuous-valued treatment, i.e., $\mathcal{A}_{S1} = \{0\}$. In S2, we consider two continuous-valued treatment options $\mathcal{A}_{S2} = \{0, 1\}$. Finally, we investigate DoseSurv’s performance in a setting with one non-dosage treatment option $\mathcal{A}_{S3}^{\text{cat}} = \{0\}$ and one continuous-valued treatment option $\mathcal{A}_{S3}^{\text{cont}} = \{1\}$. Extensive descriptions of the data generation process are provided in Appendix A.1. Our dosage assignment mechanism includes a parameter η that controls the level of dosage selection bias in the observed policy. When $\eta = 0$, the dosage is sampled from a uniform distribution between $[0, 1]$. As η increases, dosage selection becomes increasingly confounded. Throughout all experiments, we simulate dosage assignment for $\eta \in \{0, 1, 2, 3, 4\}$. In addition, we simulate a fixed treatment selection bias for scenarios S2 and S3, which comprise multiple treatment options, and employ an informed censoring process, which introduces an additional censoring bias over time.

5.1.2 Ablations on DoseSurv Design

To examine the impact of certain design elements of DoseSurv, we conduct ablation studies. We compare the standard DoseSurv model (5 RBFs with layer-specific, learnable centers initialized at $\{0, 0.25, 0.5, 0.75, 1\}$) with 5 different variants: The first ablation (A1) corresponds to the standard DoseSurv model, except with fixed (non-learnable) RBF centers. For the second variant (A2), we replace the RBF basis with a truncated polynomial (power) basis

$$\psi_{l,k}^{(a)}(q) = \begin{cases} q^k & \text{for } k = 0, \dots, d \\ \max(q - c_{l,k-d}^{(a)}, 0)^d & \text{for } k = d + 1, \dots, d + \kappa \end{cases} \quad (10)$$

to construct splines of degree $d = 2$, with $\kappa = 5$ learnable knots, $c_{l,k-d}^{(a)}$, initialized at $\{0, 0.25, 0.5, 0.75, 0.95\}$. This results in a total of $K = 8$ basis functions, and thus higher complexity than the standard DoseSurv model. The third ablation (A3) corresponds to the standard DoseSurv model, but with only three basis functions (3 centers initialized at $\{0.2, 0.5, 0.8\}$). For ablation A4, we do not add q as an additional feature to the shared representation $\Phi(\mathbf{x})$, incorporating it only implicitly by modeling the network parameters using the RBFs that depend on q . In the final ablation (A5), we employ the varying-coefficient network design proposed by Nie et al. [51] for non-TTE data. This design features a truncated polynomial basis of degree $d = 2$ with $\kappa = 2$ non-learnable knots at $\{1/3, 2/3\}$, and omits q as an additional input feature. This results in $K = 5$ basis functions, yielding a complexity comparable to our standard DoseSurv model.

5.1.3 Real-world Dataset

Additionally, we evaluate DoseSurv on the Twins dataset, a well-established benchmark in the HTE literature for binary treatments. This dataset provides survival times for twins born in the United States between 1989 and 1991 [3]. The original use of this dataset focuses on binary treatment assignment—classifying the heavier twin as "treated" and the lighter twin as "control" [46, 70]. First, we keep this notion and validate DoseSurv in a binary treatment scenario, without dosages, similar to [14]. Second, we extend the scope of the dataset by replacing the binary treatment indicator with the actual birth weight, treating birth weight as a continuous "dosage". This reinterpretation allows the dataset to represent two potential outcomes for each sample exposed to two distinct dosages. Details on the dataset and the implemented censoring mechanism are provided in Appendix A.2.

5.1.4 Baselines

We benchmark DoseSurv against well-established and state-of-the-art ML models for TTE analysis:

- **DeepSurv** [37]: A continuous-time survival model that extends the Cox proportional hazards model by incorporating NNs while retaining the assumption of proportional hazards.
- **DeepHit** [45]: A discrete-time NN method modeling the event-time distribution, trained with composite likelihood and ranking losses; permits non-proportional hazards.
- **RSF (Random Survival Forest)** [32]: A non-parametric ML ensemble of decision trees (bootstrap sampling, log-rank splits) for right-censored survival data.

- **NSC (Neural Survival Clustering)** [33]: A method combining NNs and clustering to model TTE data by grouping samples with similar survival patterns and learning hazard functions per cluster.

By default, we implement all baselines as single models (S-learner) with treatment and dosage information as additional covariates. For scenarios S2 and S3 where multiple treatment options are available, we also compare against treatment-specific versions (T-learner) of the baseline models described above—trained separately per arm with the corresponding dosage included as an input feature—and refer to them as **DeepSurv-T**, **DeepHit-T**, **RSF-T** and **NSC-T**. Moreover, we consider a version of DoseSurv without IPM regularization, **DoseSurv (no IPM)**. For the binary treatment scenario on the Twins dataset, we additionally compare DoseSurv against existing binary HTE survival models, namely **SurvITE** [14] and **BITES** [58].

5.1.5 Performance Measures

For simulated data, we have access to the ground-truth individual dose-response relationships for every treatment a and dosage $q \in [0, 1]$. We can therefore quantify how closely the estimated patient-specific dose-response curves match the true underlying curves across dosages and treatments using the mean integrated squared error (MISE)⁶ [60, 9], with RMST as the outcome of interest

$$\text{MISE} = \frac{1}{|\mathcal{A}|} \frac{1}{N} \sum_{a \in \mathcal{A}} \sum_{i=1}^N \int_{q=0}^1 (\hat{t}_{a,q}^{\text{rmst}}(\mathbf{x}_i) - t_{a,q}^{\text{rmst}}(\mathbf{x}_i))^2 dq. \quad (11)$$

For continuous-valued treatments $a \in \mathcal{A}^{\text{cont}}$, we also evaluate the mean dosage policy error (DPE) [60], i.e., the mean squared error between the ground-truth outcome under the estimated optimal dosage $\hat{q}_{i,a}^* = \arg \max_{q \in [0,1]} \hat{t}_{a,q}^{\text{rmst}}(\mathbf{x}_i)$ and the true optimal dosage $q_{i,a}^* = \arg \max_{q \in [0,1]} t_{a,q}^{\text{rmst}}(\mathbf{x}_i)$, maximizing RMST. This is given by

$$\text{DPE} = \frac{1}{|\mathcal{A}^{\text{cont}}|} \frac{1}{N} \sum_{a \in \mathcal{A}^{\text{cont}}} \sum_{i=1}^N (t_{a,\hat{q}_{i,a}^*}^{\text{rmst}}(\mathbf{x}_i) - t_{a,q_{i,a}^*}^{\text{rmst}}(\mathbf{x}_i))^2. \quad (12)$$

The performance measures described above require knowledge of the outcome generation mechanism and are thus impractical for real-world applications. Therefore, we also provide common metrics in survival analysis based on observed (factual) outcomes: the time-dependent C-index (C^{td}), the integrated Brier score (IBS), and the integrated negative binomial log-likelihood (INBLL). Details on these metrics and corresponding model performance are provided in Appendix B.

In the Twins dataset adapted for continuous-valued treatments, for each sample i , we have access to the event times $T_0^{(i)} = T_{0,q_i^{(0)}}^{(i)}$ and $T_1^{(i)} = T_{0,q_i^{(1)}}^{(i)}$ under the lower dosage (birth weight) $q_i^{(0)}$, and the higher dosage $q_i^{(1)}$ of the same treatment $\mathcal{A} = \{0\}$, but not across the whole dosage range. Instead of computing MISE, we therefore provide the MSE between the observed and the estimated HTE

$$\text{MSE} = \frac{1}{N} \sum_i (\hat{\nu}_{\text{rmst},i} - \nu_{\text{rmst},i})^2, \quad (13)$$

where $\nu_{\text{rmst},i} = \min(T_{0,q_i^{(1)}}^{(i)}, t_m) - \min(T_{0,q_i^{(0)}}^{(i)}, t_m)$ and $\hat{\nu}_{\text{rmst},i} = \hat{t}_{0,q_i^{(1)}}^{\text{rmst}}(\mathbf{x}_i) - \hat{t}_{0,q_i^{(0)}}^{\text{rmst}}(\mathbf{x}_i)$ are the observed and estimated HTEs in terms of the RMST. Analogously, we compute the MSE between estimated and observed HTEs for the non-dosage (binary) treatment case, $\mathcal{A} = \mathcal{A}^{\text{cat}} = \{0, 1\}$, where $T_1^{(i)}$ represents the survival time of the heavier twin (treated) and $T_0^{(i)}$ of the lighter twin (control).

5.2 Results and Discussion

5.2.1 Main Results

Fig. 2 shows model performance across five runs for scenarios S1–S3 under varying levels of dosage selection bias ($\eta \in \{0, 1, 2, 3, 4\}$). In all experiments, either DoseSurv or its derivative without IPM regularization consistently achieved the best performance in terms of both MISE and DPE,

⁶For non-dosage treatments $a \in \mathcal{A}^{\text{cat}}$, $t_{a,q}^{\text{rmst}}(\mathbf{x})$ is constant over q , reducing MISE to a regular MSE between predicted and ground-truth RMST.

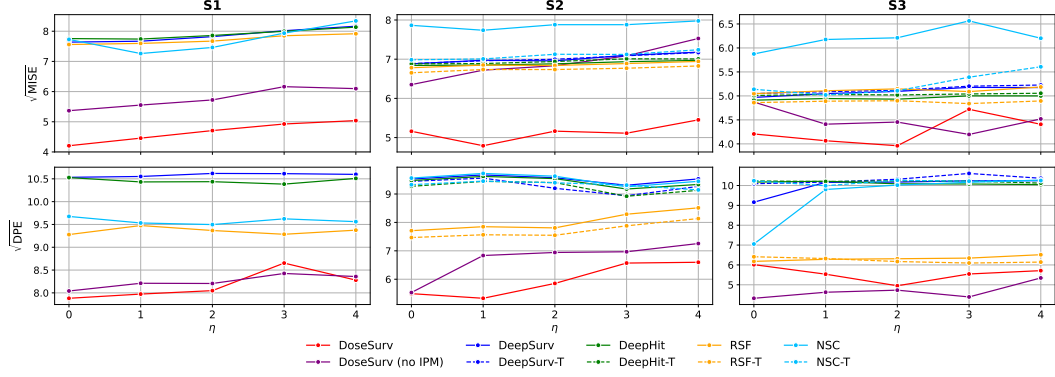


Figure 2: Model performance for scenarios S1–S3 under different levels of dosage-assignment bias, η , averaged across 5 runs. Smaller values for $\sqrt{\text{MISE}}$ and $\sqrt{\text{DPE}}$ indicate better performance.

outperforming all baseline models. As expected, performance generally declined slightly with increasing η for both metrics. Notably, IPM regularization substantially improved performance in most scenarios, particularly in S2, which involves two continuous-valued treatments, and primarily in terms of MISE. IPM regularization also yielded performance improvements even in the absence of dosage-selection bias ($\eta = 0$) and treatment-selection bias (S1). We attribute this to its ability to mitigate other covariate shifts inherent in TTE data, such as event- or censoring-induced shifts.

The baseline models did not exhibit a clear performance trend. Among them, RSF and RSF-T generally achieved the lowest DPE. NSC performed comparably well in terms of DPE for S1, but showed the worst MISE performance in S2 and S3. DeepSurv(-T), DeepHit(-T), and NSC-T consistently demonstrated similar performance across all metrics and experiments.

Table 1 summarizes the performance on the real-world Twins dataset. Modeling the problem as a continuous treatment scenario and incorporating dosage information substantially improved performance across all models, highlighting the importance of continuous treatment modeling in survival analysis. DoseSurv demonstrated the best performance in estimating the HTE in both binary and continuous treatment settings, with substantial performance improvement observed in the continuous case. Among the baselines, DeepHit achieved the highest performance in the continuous setting, while RSF and DeepHit performed equally well in the binary setting.

Table 1: Model performance on the Twins dataset across 5 runs (mean $\pm$ 95% CI). For baselines, we also show the relative difference compared to DoseSurv.

Methods	Twins (continuous)	Twins (binary)
	$\sqrt{\text{MSE}}$	$\sqrt{\text{MSE}}$
DeepSurv	14.461 $\pm$ 0.217 (+15.7%)	16.235 $\pm$ 0.151 (+4.1%)
DeepHit	13.272 $\pm$ 0.203 (+6.2%)	15.974 $\pm$ 0.090 (+2.4%)
RSF	14.230 $\pm$ 0.092 (+13.8%)	15.923 $\pm$ 0.022 (+2.1%)
NSC	14.288 $\pm$ 0.554 (+14.3%)	16.694 $\pm$ 0.123 (+7.0%)
BITES	—	16.177 $\pm$ 0.161 (+3.7%)
SurvITE	—	19.599 $\pm$ 4.725 (+25.6%)
DoseSurv	12.501$\pm$0.063	15.601$\pm$0.531

5.2.2 Ablation Study Results

Figure 3 illustrates DoseSurv’s performance for scenario S2 ($\eta = 1$), compared to various ablations (A1–A5), described in Section 5.1.2. The proposed design consistently achieved the best performance across all metrics compared to other variants. In particular, each individual design choice yielded a performance improvement over a DoseSurv implementation based on a traditional varying-coefficient network design. Moreover, the use of RBFs results in fewer basis functions and coefficients, $\beta_{j,l,k}^{(a)}$, and thus lower model complexity, for the same number of centers/knots.

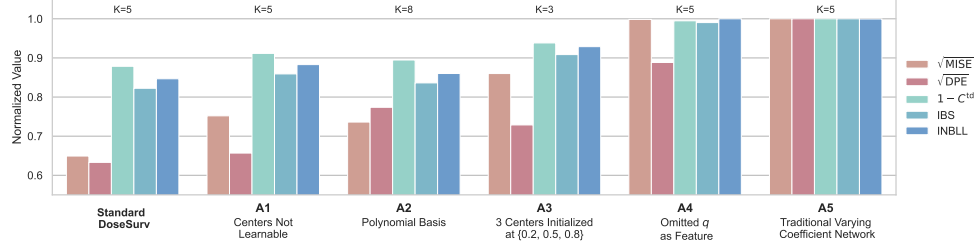


Figure 3: Impact of design choices (A1-A5) on the performance of DoseSurv in scenario S2 ($\eta = 1$) averaged across 5 runs. All other hyperparameters were held constant. Performance metrics were normalized such that the highest mean value across all configurations equals 1. Lower values imply better performance. We indicate model complexity in terms of the number of basis functions, K .

6 Conclusion

In this paper, we presented an RBF-based varying-coefficient network for estimating HTEs from TTE data in single- and multi-arm treatment regimens with continuous-valued (dosage-dependent) and non-dosage treatment options. We demonstrated in various experiments the benefit of the varying-coefficient design in TTE settings with continuous treatments. Furthermore, we demonstrated significant performance improvements by incorporating an IPM regularization scheme tailored for these settings, effectively mitigating covariate shifts caused by confounding factors and the nature of TTE data.

Limitations Like all HTE methods, our model relies on certain assumptions. In practice, leveraging expert and domain knowledge helps assess these and understand the impact of possible violations. Another challenge is tuning hyperparameters—especially the IPM regularization strength γ —under covariate shift. One way to address this is to select the largest γ that preserves discrimination and calibration (e.g., via INBLL) on the validation set, balancing regularization and predictive power [14]. Finally, as with most deep learning models, performance benefits from sufficient training data; our sample size ablation in Appendix C.1 indicates degradation with small training sets.

Broader Impact A promising future application of models like those explored in this work is adaptive clinical trials, where integrating observational TTE data and leveraging ML can optimize trial designs. This approach could enable dynamic adjustments in patient recruitment, treatment, and dosage options, ultimately improving trial outcomes and patient care. However, analyses must contend with covariate shifts—including treatment- and dosage-selection bias in observational settings and event- or censoring-induced shifts inherent to survival data—which, if unaddressed, can bias effect estimates and hinder optimal care.

Acknowledgments

MG is supported by the EPSRC Centre for Doctoral Training in Health Data Science (EP/S02428X/1). CY is supported by a UKRI Turing AI Acceleration Fellowship (EP/V023233/1) and received additional funding through EPSRC grant EP/Y018192/1. TZ is supported by the Royal Academy of Engineering under the Research Fellowship scheme.

References

- [1] A. M. Alaa and M. van der Schaar. Bayesian inference of individualized treatment effects using multi-task gaussian processes. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
- [2] A. M. Alaa, M. Weisz, and M. van der Schaar. Deep counterfactual networks with propensity-dropout, 2017.

- [3] D. Almond, K. Y. Chay, and D. S. Lee. The costs of low birth weight. *The Quarterly Journal of Economics*, 120(3):1031–1083, Aug. 2005.
- [4] L. Antolini, P. Boracchi, and E. Biganzoli. A time-dependent discrimination index for survival data. *Statistics in Medicine*, 24(24):3927–3944, Nov. 2005.
- [5] S. Assaad, S. Zeng, C. Tao, S. Datta, N. Mehta, R. Henao, F. Li, and L. Carin Duke. Counterfactual representation learning with balancing weights. In A. Banerjee and K. Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1972–1980. PMLR, 13–15 Apr 2021.
- [6] S. Athey, J. Tibshirani, and S. Wager. Generalized random forests. *The Annals of Statistics*, 47(2), Apr. 2019.
- [7] E. Bair and R. Tibshirani. Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biol.*, 2(4):E108, Apr. 2004.
- [8] A. Bellot, A. Dhir, and G. Prando. Generalization bounds and algorithms for estimating conditional average treatment effect of dosage, 2022.
- [9] I. Bica, J. Jordon, and M. van der Schaar. Estimating the effects of continuous-valued interventions using generative adversarial networks. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 16434–16445. Curran Associates, Inc., 2020.
- [10] E. Biganzoli, P. Boracchi, L. Mariani, and E. Marubini. Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach. *Statistics in Medicine*, 17(10):1169–1186, May 1998.
- [11] C. Bockel genannt Rickermann, T. Vanderschueren, J. Berrevoets, T. Verdonck, and W. Verbeke. Using representation balancing to learn conditional-average dose responses from clustered data, 2025.
- [12] P. Chapfuwa, S. Assaad, S. Zeng, M. J. Pencina, L. Carin, and R. Henao. Enabling counterfactual survival analysis with balanced representations. In *Proceedings of the Conference on Health, Inference, and Learning*, ACM CHIL ’21. ACM, Apr. 2021.
- [13] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 34(2):187–202, Jan. 1972.
- [14] A. Curth, C. Lee, and M. van der Schaar. Survite: Learning heterogeneous treatment effects from time-to-event data. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 26740–26753. Curran Associates, Inc., 2021.
- [15] A. Curth and M. van der Schaar. Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms. In A. Banerjee and K. Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1810–1818. PMLR, 13–15 Apr 2021.
- [16] A. Curth and M. van der Schaar. On inductive biases for heterogeneous treatment effect estimation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 15883–15894. Curran Associates, Inc., 2021.
- [17] M. Cuturi and A. Doucet. Fast computation of wasserstein barycenters. In E. P. Xing and T. Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 685–693, Beijing, China, 22–24 Jun 2014. PMLR.
- [18] L. R. Duska, R. Penson, J. G. Supko, D. M. Finkelstein, T. Makastorsis, J. Gallagher, K. Borden, A. Goodman, A. F. Fuller, N. Nikrui, and M. V. Seiden. A phase i study of continuous infusion doxorubicin and paclitaxel chemotherapy with granulocyte colony-stimulating factor for relapsed epithelial ovarian cancer1. *Clinical Cancer Research*, 5(6):1299–1305, 06 1999.

- [19] I. Díaz, O. Savenkov, and K. Ballman. Targeted learning ensembles for optimal individualized treatment rules with time-to-event outcomes. *Biometrika*, 105(3):723–738, May 2018.
- [20] D. Faraggi and R. Simon. A neural network model for survival data. *Statistics in Medicine*, 14(1):73–82, Jan. 1995.
- [21] M. Fiuzat, D. Wojdyla, D. Kitzman, J. Fleg, S. J. Keteyian, W. E. Kraus, I. L. Piña, D. Whellan, and C. M. O’Connor. Relationship of beta-blocker dose with outcomes in ambulatory heart failure patients with systolic dysfunction. *Journal of the American College of Cardiology*, 60(3):208–215, July 2012.
- [22] J. A. Foekens, H. A. Peters, M. P. Look, H. Portengen, M. Schmitt, M. D. Kramer, N. Brünnner, F. Jänicke, M. E. M.-v. Gelder, S. C. Henzen-Logmans, W. L. J. van Putten, and J. G. M. Klijn. The urokinase system of plasminogen activation and prognosis in 2780 breast cancer patients1. *Cancer Research*, 60(3):636–643, 02 2000.
- [23] J. C. Foster, J. M. Taylor, and S. J. Ruberg. Subgroup identification from randomized clinical trial data. *Statistics in Medicine*, 30(24):2867–2880, Aug. 2011.
- [24] M. F. Gensheimer and B. Narasimhan. A scalable discrete-time survival model for neural networks. *PeerJ*, 7:e6257, Jan. 2019.
- [25] T. A. Gerds and M. Schumacher. Consistent estimation of the expected brier score in general survival models with right-censored event times. *Biometrical Journal*, 48(6):1029–1040, Dec. 2006.
- [26] E. Graf, C. Schmoor, W. Sauerbrei, and M. Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18(17–18):2529–2545, Sept. 1999.
- [27] N. Hassanpour and R. Greiner. Counterfactual regression with importance sampling weights. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5880–5887. International Joint Conferences on Artificial Intelligence Organization, 7 2019.
- [28] T. Hastie and R. Tibshirani. Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(4):757–796, 1993.
- [29] N. C. Henderson, T. A. Louis, G. L. Rosner, and R. Varadhan. Individualized treatment effects with censored data via fully nonparametric bayesian accelerated failure time models. *Biostatistics*, 21(1):50–68, July 2018.
- [30] G. W. Imbens. The role of the propensity score in estimating dose-response functions. *Biometrika*, 87(3):706–710, 2000.
- [31] G. W. Imbens and D. B. Rubin. *Rubin Causal Model*, page 229–241. Palgrave Macmillan UK, 2010.
- [32] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer. Random survival forests. *The Annals of Applied Statistics*, 2(3), Sept. 2008.
- [33] V. Jeanselme, B. Tom, and J. Barrett. Neural survival clustering: Non-parametric mixture of neural networks for survival clustering. In G. Flores, G. H. Chen, T. Pollard, J. C. Ho, and T. Naumann, editors, *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 92–102. PMLR, 07–08 Apr 2022.
- [34] F. Johansson, U. Shalit, and D. Sontag. Learning representations for counterfactual inference. In M. F. Balcan and K. Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 3020–3029, New York, New York, USA, 20–22 Jun 2016. PMLR.
- [35] J. D. Kalbfleisch and R. L. Prentice. *The Statistical Analysis of Failure Time Data*. Wiley, Aug. 2002.

- [36] U. Kampmann, S. Knorr, J. Fuglsang, and P. Ovesen. Determinants of maternal insulin resistance during pregnancy: An updated overview. *Journal of Diabetes Research*, 2019:1–9, Nov. 2019.
- [37] J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger. DeepSurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1), Feb. 2018.
- [38] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, San Diego, CA, USA, 2015.
- [39] N. Kreif, K. DiazOrdaz, R. Moreno-Serra, A. Mirelman, T. Hidayat, and M. Suhrcke. Estimating heterogeneous policy impacts using causal machine learning: a case study of health insurance reform in Indonesia. *Health Services and Outcomes Research Methodology*, 22(2):192–227, Nov. 2021.
- [40] S. R. Künzel, J. S. Sekhon, P. J. Bickel, and B. Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, Feb. 2019.
- [41] H. Kvamme and Ø. Borgan. Continuous and discrete-time survival prediction with neural networks. *Lifetime Data Analysis*, 27(4):710–736, Oct. 2021.
- [42] H. Kvamme and Ø. Borgan. The brier score under administrative censoring: Problems and a solution. *Journal of Machine Learning Research*, 24(2):1–26, 2023.
- [43] H. Kvamme, Ø. Borgan, and I. Scheel. Time-to-event prediction with neural networks and cox regression. *Journal of Machine Learning Research*, 20(129):1–30, 2019.
- [44] H. Langen and M. Huber. How causal machine learning can leverage marketing strategies: Assessing and improving the performance of a coupon campaign. *PLOS ONE*, 18(1):e0278937, Jan. 2023.
- [45] C. Lee, W. Zame, J. Yoon, and M. Van der Schaar. DeepHit: A deep learning approach to survival analysis with competing risks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018.
- [46] C. Louizos, U. Shalit, J. M. Mooij, D. Sontag, R. Zemel, and M. Welling. Causal effect inference with deep latent-variable models. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
- [47] M. Lu, S. Sadiq, D. J. Feaster, and H. Ishwaran. Estimating individual treatment effect in observational data using random forest methods. *Journal of Computational and Graphical Statistics*, 27(1):209–219, Jan. 2018.
- [48] A. D. Meid, C. Ruff, L. Wirbka, F. Stoll, H. M. Seidling, A. Groll, and W. E. Haefeli. Using the causal inference framework to support individualized drug treatment decisions based on observational healthcare data. *Clinical Epidemiology*, Volume 12:1223–1234, Nov. 2020.
- [49] M. Mesinovic, P. Watkinson, and T. Zhu. Dysurv: dynamic deep learning model for survival analysis with conditional variational inference. *Journal of the American Medical Informatics Association*, page ocae271, 11 2024.
- [50] C. Nagpal, M. Goswami, K. Dufendach, and A. Dubrawski. Counterfactual phenotyping with censored time-to-events. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD ’22. ACM, Aug. 2022.
- [51] L. Nie, M. Ye, Q. Liu, and D. Nicolae. VCNNet and functional targeted regularization for learning causal effects of continuous treatments. In *International Conference on Learning Representations*, 2021.
- [52] J. Park and I. W. Sandberg. Universal approximation using radial-basis-function networks. *Neural Computation*, 3(2):246–257, June 1991.

- [53] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Sept. 2009.
- [54] S. Pölsterl. scikit-survival: A library for time-to-event analysis built on top of scikit-learn. *Journal of Machine Learning Research*, 21(212):1–6, 2020.
- [55] R. Ranganath, A. Perotte, N. Elhadad, and D. Blei. Deep survival analysis. In F. Doshi-Velez, J. Fackler, D. Kale, B. Wallace, and J. Wiens, editors, *Proceedings of the 1st Machine Learning for Healthcare Conference*, volume 56 of *Proceedings of Machine Learning Research*, pages 101–114, Northeastern University, Boston, MA, USA, 18–19 Aug 2016. PMLR.
- [56] K. Ren, J. Qin, L. Zheng, Z. Yang, W. Zhang, L. Qiu, and Y. Yu. Deep recurrent survival analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):4798–4805, July 2019.
- [57] P. Royston and D. G. Altman. External validation of a cox prognostic model: principles and methods. *BMC Medical Research Methodology*, 13(1), Mar. 2013.
- [58] S. Schrod, A. Schäfer, S. Solbrig, R. Lohmayer, W. Gronwald, P. J. Oefner, T. Beißbarth, R. Spang, H. U. Zacharias, and M. Altenbuchinger. Bites: balanced individual treatment effect for survival data. *Bioinformatics*, 38(Supplement_1):i60–i67, June 2022.
- [59] M. Schumacher, G. Bastert, H. Bojar, K. Hübner, M. Olschewski, W. Sauerbrei, C. Schmoor, C. Beyerle, R. L. Neumann, and H. F. Rauschecker. Randomized 2 x 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. german breast cancer study group. *Journal of Clinical Oncology*, 12(10):2086–2093, Oct. 1994.
- [60] P. Schwab, L. Linhardt, S. Bauer, J. M. Buhmann, and W. Karlen. Learning counterfactual representations for estimating individual dose-response curves. In *Proceedings of the 34th Conference on Artificial Intelligence (AAAI)*, volume 34, pages 5612–5619. AAAI Press, Feb. 2020. AAAI Technical Track: Machine Learning.
- [61] J. Schweisthal, D. Frauen, V. Melnychuk, and S. Feuerriegel. Reliable off-policy learning for dosage combinations. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 67900–67924. Curran Associates, Inc., 2023.
- [62] U. Shalit, F. D. Johansson, and D. Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3076–3085. PMLR, 06–11 Aug 2017.
- [63] C. Shi, D. Blei, and V. Veitch. Adapting neural networks for the estimation of treatment effects. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.
- [64] O. M. Stitelman and M. J. van der Laan. Collaborative targeted maximum likelihood for time to event data. *The International Journal of Biostatistics*, 6(1), Jan. 2010.
- [65] O. M. Stitelman, C. W. Wester, V. De Gruttola, and M. J. van der Laan. Targeted maximum likelihood estimation of effect modification parameters in survival analysis. *The International Journal of Biostatistics*, 7(1):1–34, Jan. 2011.
- [66] K. Suresh, C. Severn, and D. Ghosh. Survival prediction models: an introduction to discrete-time modeling. *BMC Medical Research Methodology*, 22(1), July 2022.
- [67] S. Tabib and D. Larocque. Non-parametric individual treatment effect estimation for survival data with random forests. *Bioinformatics*, 36(2):629–636, Aug. 2019.
- [68] C. Villani. *The Wasserstein distances*, page 93–111. Springer Berlin Heidelberg, 2009.
- [69] S. Wager and S. Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, June 2018.

- [70] J. Yoon, J. Jordon, and M. van der Schaar. GANITE: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations*, 2018.
- [71] W. Zhang, T. D. Le, L. Liu, Z.-H. Zhou, and J. Li. Mining heterogeneous causal effects for personalized cancer treatment. *Bioinformatics*, 33(15):2372–2378, Mar. 2017.
- [72] Y.-F. Zhang, H. Zhang, Z. C. Lipton, L. E. Li, and E. P. Xing. Exploring transformer backbones for heterogeneous treatment effect estimation, 2022.

Appendix

A Data Description

A.1 Synthetic Data Generation

We generate covariates from a correlated multivariate normal distribution, $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \Sigma)$ with $\Sigma = (1 - \rho)\mathbf{I} + \rho\mathbf{1}\mathbf{1}^\top$ and $\rho = 0.2$. We generate 20,000 samples with 40 covariates for scenario S1 and 20 covariates for S2 and S3. To simulate event times, we randomly sample a number of parameter vectors relevant to define the event-processes of each scenario, $\mathbf{u}_j^{S1}, \mathbf{u}_j^{S2}, \mathbf{u}_j^{S3} \sim \mathcal{N}(0, \mathbf{I}^2)$ and normalize them $\mathbf{v}_j^S = \mathbf{u}_j^S / \|\mathbf{u}_j^S\|$, where $\|\cdot\|$ is the Euclidean norm, and $S \in \{S1, S2, S3\}$. In scenarios S2 and S3, involving two treatment options, we aim to simulate an event process that reflects shared characteristics across treatments (e.g., common risk factors such as age) while also incorporating treatment-specific properties in the event process. We denote shared parameter vectors as $\mathbf{v}_j^S$ and treatment-specific parameter vectors as $\tilde{\mathbf{v}}_j^{S,a}$. Finally, for the three scenarios S1, S2, S3, we use the following event processes, where $\sigma(\cdot)$ is the sigmoid function:

S1 (1 continuous-valued treatment option)
$\lambda_{0,q}(t \mathbf{x}) = 0.25\sigma(\mathbf{v}_1^{S1}\mathbf{x})\sigma(-2\mathbf{v}_2^{S1}\mathbf{x}\sin(2\pi\mathbf{v}_3^{S1}\mathbf{x}(q+1)t/t_m) - 3.5\sin(6\pi q\sigma(\mathbf{v}_4^{S1}\mathbf{x} + 2/3) + \mathbf{v}_5^{S1}\mathbf{x}) - \text{sgn}(\mathbf{v}_6^{S1}\mathbf{x})q^2)$
S2 (2 continuous-valued treatment options)
$\lambda_{0,q}(t \mathbf{x}) = \begin{cases} 0.1\sigma(\mathbf{v}_1^{S2}\mathbf{x}), & \text{if } t < 5 \\ 0.5\sigma(\mathbf{v}_1^{S2}\mathbf{x})\sigma\left(5\sin(5\pi q + \tilde{\mathbf{v}}_1^{S2,0}\mathbf{x}) + 2\tilde{\mathbf{v}}_2^{S2,0}\mathbf{x}q^2\right), & \text{if } t \geq 5 \end{cases}$ $\lambda_{1,q}(t \mathbf{x}) = \begin{cases} 0.1\sigma(\mathbf{v}_1^{S2}\mathbf{x}), & \text{if } t < 5 \\ 0.5\sigma(\mathbf{v}_1^{S2}\mathbf{x})\sigma\left(5\text{sgn}(\tilde{\mathbf{v}}_1^{S2,1}\mathbf{x})\sin(4\pi q + \tilde{\mathbf{v}}_2^{S2,1}\mathbf{x}) + \tilde{\mathbf{v}}_1^{S2,1}\mathbf{x}q^2\right), & \text{if } t \geq 5 \end{cases}$
S3 (1 non-dosage and 1 continuous-valued treatment option)
$\lambda_0(t \mathbf{x}) = \begin{cases} 0.1\sigma(\mathbf{v}_1^{S3}\mathbf{x}), & \text{if } t < 5 \\ 0.2\sigma(3\tilde{\mathbf{v}}_1^{S3,0}\mathbf{x} - 1), & \text{if } t \geq 5 \end{cases}$ $\lambda_{1,q}(t \mathbf{x}) = \begin{cases} 0.1\sigma(\mathbf{v}_1^{S3}\mathbf{x}), & \text{if } t < 5 \\ 0.5\sigma(\mathbf{v}_1^{S3}\mathbf{x})\sigma\left(5\sin(5\pi q + \tilde{\mathbf{v}}_1^{S3,1}\mathbf{x}) + 2\tilde{\mathbf{v}}_2^{S3,1}\mathbf{x}q^2\right), & \text{if } t \geq 5 \end{cases}$

Fig. A.1 (A) shows simulated ground-truth survival curves for scenario S1 for 6 different samples and dosages $q = 0.2, 0.5, 0.8$. Fig. A.1 (B) depicts the corresponding dose-response relationships, where we define the outcome of interest as the RMST. We define the covariate-dependent, informative, censoring process as

$$\lambda_{\text{cens}}^S(t|\mathbf{x}) = 0.02\sigma(\mathbf{v}_{\text{cens}}^S\mathbf{x}),$$

where $\mathbf{v}_{\text{cens}}^S$ is another randomly sampled and normalized parameter vector, introducing an additional censoring bias, and therefore covariate shift over time. Furthermore, we assume administrative censoring at $t_m = 30$. To assign dosages, we follow an approach similar to [9] and [61]. More specifically, for each continuous treatment option, a , we sample the dosages from a Beta distribution

$$q_a \sim \text{Beta}(\tilde{\eta}, b_{q_a^*})$$

with $b_{q_a^*} = (\tilde{\eta} - 1)/q_a^* + 2 - \tilde{\eta}$, where the mode of the Beta distribution is given by the true optimal dosage, q_a^* , maximizing the RMST for the respective patient under treatment a .⁷ This mechanism mimics a realistic medical scenario in which the dosage assignment is informed by the true optimal

⁷To maintain symmetry, we sample $q_a \sim \text{Beta}(b_{q_a^*}, \tilde{\eta})$ with $b_{q_a^*} = (\tilde{\eta} - 1)/(1 - q_a^*) + 2 - \tilde{\eta}$ if $q_a^* < 0.5$.

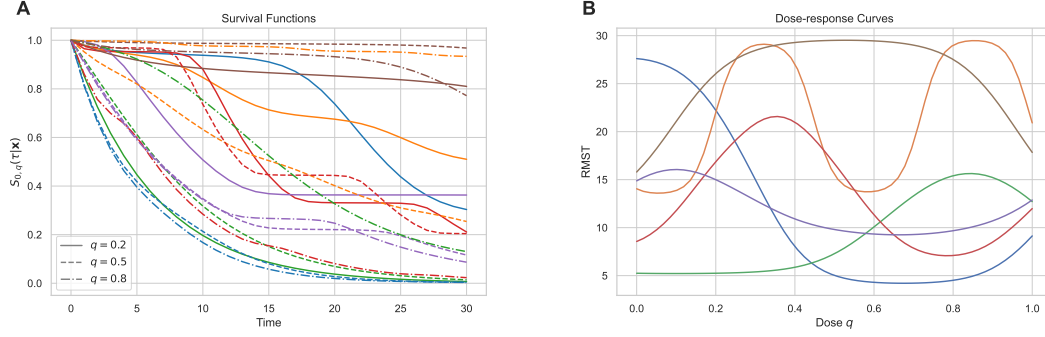


Figure A.1: Simulated (continuous) ground-truth survival curves for scenario S1 for 6 different samples and dosages $q = 0.2, 0.5, 0.8$ (A). Corresponding dose-response relationships with RMST as the outcome of interest (B).

dosage. The parameter $\eta = \tilde{\eta} - 1$ controls the dosage assignment, i.e., how much influence the covariates have on the dosage assignment. With $\eta = 0$, the dosage is sampled from a uniform distribution, while the dosage assignment becomes increasingly confounded towards the optimal dosage as η increases.

To simulate confounded treatment assignment in scenarios S2 and S3, we design a mechanism that links the observed treatment type a with the individual patient covariates. For this, we sample M parameter vectors $\mathbf{u}_a \sim \mathcal{N}(0, \mathbf{I}^2)$ and normalize them $\mathbf{v}_a = \mathbf{u}_a / \|\mathbf{u}_a\|$. Then, we standardize the covariates $\mathbf{x}$ and compute the probabilities for observing treatment a

$$p_a = \frac{\exp(\mathbf{x}\mathbf{v}_a)}{\sum_{\alpha \in \mathcal{A}} \exp(\mathbf{x}\mathbf{v}_\alpha)}.$$

Next, we compute the modulated probabilities

$$\tilde{p}_a = \frac{(p_a)^\xi}{\sum_{\alpha \in \mathcal{A}} (p_\alpha)^\xi},$$

where ξ controls the treatment selection bias or confounding, i.e., how much influence the covariates have on the treatment decision (with $\xi = 0$ indicating a random treatment decision). For both scenarios S2 and S3 we choose $\xi = 1.5$ and sample the treatment from the multinomial distribution

$$a \sim \text{Multinomial}(\{\tilde{p}_\alpha\}_{\alpha \in \mathcal{A}}),$$

suitable for any number of treatment options. Figure A.2 illustrates the covariate shift in two selected covariates, $x_{11} + x_{19}$, between the two different treatment groups in scenario S2. The histograms represent the at-risk populations at different time points ($t = 0, 10, 20$) under different values of

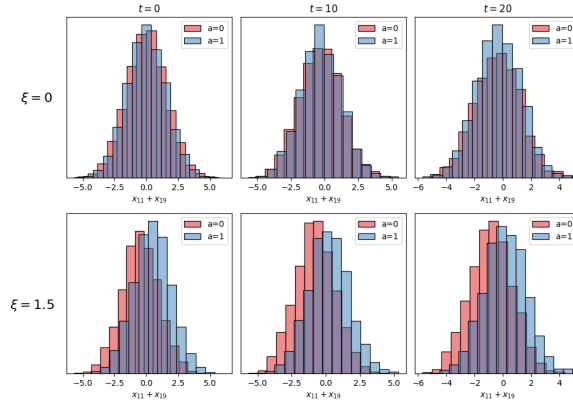


Figure A.2: Histograms of $x_{11} + x_{19}$ for at-risk populations across time for different treatments, $a = 0$ and $a = 1$, in scenario S2. Covariate shifts arise with increasing time t and bias parameter ξ .

the treatment selection bias parameter ($\xi = 0, 1.5$). As ξ increases, the covariate shifts become more pronounced. Additionally, shifts evolve over time due to differing event probabilities between treatment groups [14].

Across all scenarios, we employ 17,000 samples for training, 1,000 samples for validation, and 2,000 samples for testing.

A.2 Twins Data

The Twins data is sourced from all US births between 1989–1991 [3], where we extract data on twins only, similar to [46, 70, 14]. Moreover, we select twins weighing less than 1,100 grams, extract 45 features predominantly related to parental factors, pregnancy, and birth, and include only samples with no missing features. We analyze 30-day survival, retaining twin pairs with no death after day 30; twins without a recorded death are treated as administratively censored at day 30. Moreover, we keep the original time resolution of 1 day as the length of the time intervals ($t_m = 30$). We create a semi-synthetic observational TTE dataset by selecting one twin from each pair. For this, we use a similar mechanism as previously for the treatment assignment in the simulated scenarios S2 and S3: First, we randomly sample two parameter vectors $\mathbf{u}_0^{\text{twin}}, \mathbf{u}_1^{\text{twin}} \sim \mathcal{N}(0, 1^2)$ and normalize them $\mathbf{v}_0^{\text{twin}} = \mathbf{u}_0^{\text{twin}} / \|\mathbf{u}_0^{\text{twin}}\|, \mathbf{v}_1^{\text{twin}} = \mathbf{u}_1^{\text{twin}} / \|\mathbf{u}_1^{\text{twin}}\|$. After standardizing the covariates $\mathbf{x}$, we define the probability for observing the lower-weight twin or higher-weight twin as

$$p_a = \frac{\exp(\mathbf{x} \mathbf{v}_a^{\text{twin}})}{\sum_{\alpha \in \{0,1\}} \exp(\mathbf{x} \mathbf{v}_\alpha^{\text{twin}})}.$$

Next, we compute the modulated probabilities

$$\tilde{p}_a = \frac{(p_a)^\xi}{\sum_{\alpha \in \{0,1\}} ((p_\alpha)^\xi)},$$

where we choose $\xi = 20$. We apply a synthetic censoring mechanism inspired by Schrod et al. [58] to introduce early censoring into samples which are not already administratively censored at day 30. As previously, we generate a normalized random vector $\mathbf{v}_{\text{cens}}^{\text{twin}}$ and compute a censoring score s_i^{cens} for each patient i , by computing the dot product with the standardized covariates: $s_i^{\text{cens}} = \mathbf{v}_{\text{cens}}^{\text{twin}} \mathbf{x}_i$. After standardizing s_i^{cens} across all samples, we censor all remaining samples where $s_i^{\text{cens}} > 0$ (i.e., $\sim 50\%$ of the remaining samples) at a random fraction $f_i \sim U(0, 1)$ of the sample's true observed survival time. Note that following this process we simulate a scenario where the independent censoring assumption is violated.

The final cohort contains 5,601 samples. For our analysis, we use 70% of samples for training. From the remaining samples, we extract 30% for validation and 70% for testing.

B Metrics Based on Observed Outcomes

B.1 Definitions

In the following, we provide definitions of the additional metrics used to measure the performance of survival models based on observed (factual) outcomes. Here we treat time as continuous and interpolate accordingly.

Time-dependent C-index (C^{td}): Following the definition from [4, 43], we compute C^{td} as

$$C^{\text{td}} = \mathbb{P}(\hat{S}(\tilde{\tau}_i | \mathbf{x}_i) < \hat{S}(\tilde{\tau}_i | \mathbf{x}_j) | \tilde{\tau}_i < \tilde{\tau}_j, \delta_i = 1).$$

This metric is *not* limited to settings where the proportional hazard assumption holds, but can be biased since only uncensored individuals are taken into account [49].

Integrated (IPCW) Brier Score (IBS): The IBS [26, 25] can be interpreted as a measure of the mean accuracy of the predicted survival probability over time, adjusting for censoring by incorporating inverse probability of censoring weights (IPCW). It is computed as

$$\text{IBS} = \frac{1}{\max(\tilde{\tau}_i)} \int_0^{\max(\tilde{\tau}_i)} \frac{1}{N} \sum_{i=1}^N \left[\frac{\hat{S}(t | \mathbf{x}_i)^2 \mathbb{1}\{\tilde{\tau}_i \leq t, \delta_i = 1\}}{\hat{G}(\tilde{\tau}_i)} + \frac{(1 - \hat{S}(t | \mathbf{x}_i))^2 \mathbb{1}\{\tilde{\tau}_i > t\}}{\hat{G}(t)} \right] dt,$$

where we denote the Kaplan-Meier estimator applied to the censoring times as $\hat{G}(t) = \mathbb{P}(C > t)$. Here, biases may arise due to dependencies between censoring distribution and covariates [42, 43].

Integrated (IPCW) Negative Binomial Log-likelihood (INBLL): The INBLL measures the performance of a survival model in terms of discrimination and calibration, averaged over time. It is calculated as [43]

$$\text{INBLL} = -\frac{1}{\max(\tilde{\tau}_i)} \int_0^{\max(\tilde{\tau}_i)} \frac{1}{N} \sum_{i=1}^N \left[\frac{\log[\hat{S}(t|\mathbf{x}_i)] \mathbb{1}\{\tilde{\tau}_i > t\}}{\hat{G}(t)} + \frac{\log[1 - \hat{S}(t|\mathbf{x}_i)] \mathbb{1}\{\tilde{\tau}_i \leq t, \delta_i = 1\}}{\hat{G}(\tilde{\tau}_i)} \right] dt.$$

We use this metric for hyperparameter optimization of the baseline models (see Section E.2).

B.2 Results on Observable Metrics

Fig. B.1 illustrates the model performance, measured by C^{td} , IBS, and INBLL, for scenarios S1–S3, averaged over five runs. The results align closely with the synthetic data metrics presented in the main paper (see Fig. 2). Overall, DoseSurv consistently outperforms other models across most metrics and scenarios, with its variant without IPM regularization typically ranking second. Among the baseline models, DeepSurv generally achieves the best performance in terms of C^{td} , IBS, and INBLL in scenario S1, while RSF-T demonstrates high performance in scenarios S2 and S3.

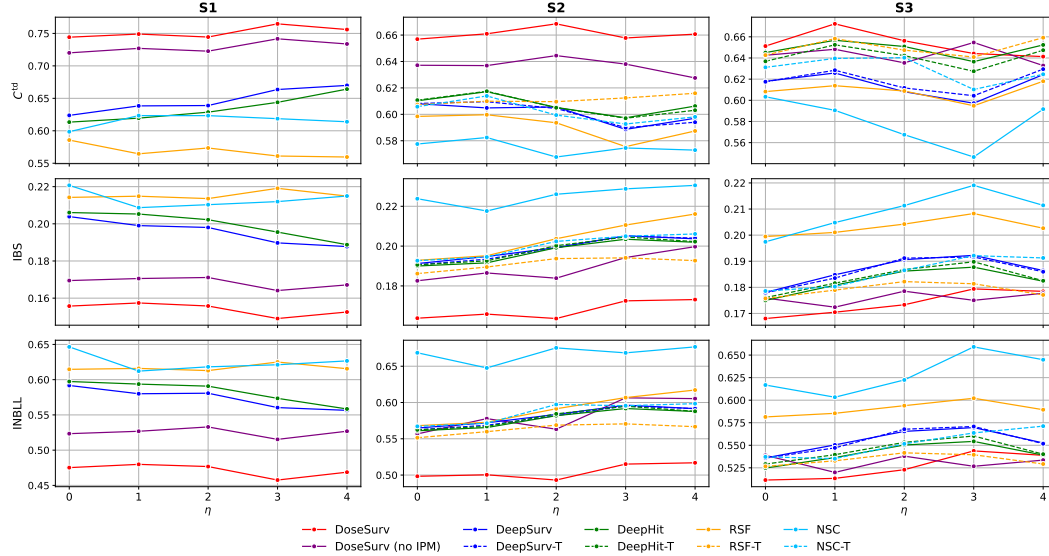


Figure B.1: Model performance for scenarios S1–S3, in terms of C^{td} , IBS and INBLL under different levels of dosage-assignment bias η , averaged across 5 runs. Greater values of C^{td} and lower values of IBS and INBLL indicate better performance.

C Additional Experiments

C.1 Ablation Study on Sample Size

We conducted an ablation study to examine the impact of training dataset size on model performance in scenario S1 ($\eta = 0$). We performed experiments using 5 different training sample sizes: 2,000, 6,000, 10,000, 14,000 and 18,000. Each training dataset was independently and randomly sampled from the combined training and validation pool. The same hyperparameters as in the main experiment for S1 were employed. For methods using early stopping (DoseSurv, DeepHit, DeepSurv, NSC), we reserved 10% of each sampled dataset for validation; RSF was trained on the full sampled set. The results are depicted in Fig. C.1. DoseSurv achieved the best performance across all sizes and metrics, except at the smallest size.

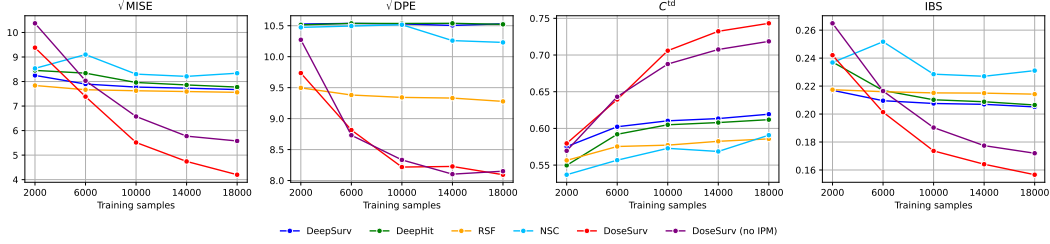


Figure C.1: Model performance across 5 runs (mean $\pm$ 95% CI) for scenario S1 ($\eta = 0$) for different training dataset sizes.

C.2 Binary Treatment Scenario: Hormone Therapy in Breast Cancer Patients

We outline a potential application of DoseSurv to optimize binary treatment decisions, specifically hormone therapy for breast cancer patients. This external validation scenario has been previously explored by [57, 37, 58]. Although binary treatment settings are not the main focus of DoseSurv, the same approach could be applied to a continuous-valued treatment problem.

Following the methodology established in the referenced literature, we use data from 1,545 breast cancer patients in the Rotterdam Tumor Bank [22]. From this observational dataset, we extract six features: age, menopausal status, number of cancerous lymph nodes, tumor size, and progesterone and estrogen receptor statuses. Since treatment assignment in the Rotterdam dataset is non-random, we use these data for model training (85% training, 15% validation).

For testing, we use data from 686 patients enrolled in a randomized controlled trial conducted by the German Breast Cancer Study Group (GBSG) [59], ensuring external validation. The same six patient characteristics are extracted from this dataset. To simplify the analysis, survival times were discretized into 30 equally-sized intervals for DoseSurv and all baseline models.

The results, summarized in Table C.1, demonstrate that DoseSurv, although designed primarily for continuous-valued treatments, can be applied effectively in binary treatment settings. In particular, DoseSurv achieved better calibration accuracy (lower IBS and INBLL) than baseline models. Furthermore, IPM regularization led to improvements in model performance.

Table C.1: Model performance on the GBSG dataset across 5 runs (mean $\pm$ 95% CI).

Methods	C^{td}	IBS	INBLL
DeepSurv	0.683 $\pm$ 0.004	0.118 $\pm$ 0.004	0.384 $\pm$ 0.012
DeepHit	0.680 $\pm$ 0.009	0.121 $\pm$ 0.022	0.394 $\pm$ 0.055
RSF	0.685 $\pm$ 0.001	0.122 $\pm$ 0.001	0.403 $\pm$ 0.002
NSC	0.672 $\pm$ 0.012	0.108 $\pm$ 0.020	0.370 $\pm$ 0.058
BITES	0.675 $\pm$ 0.013	0.115 $\pm$ 0.004	0.381 $\pm$ 0.008
SurvITE	0.665 $\pm$ 0.037	0.187 $\pm$ 0.072	0.539 $\pm$ 0.174
DoseSurv (no IPM)	0.670 $\pm$ 0.008	0.110 $\pm$ 0.005	0.356 $\pm$ 0.012
DoseSurv	0.688$\pm$0.002	0.098$\pm$0.003	0.338$\pm$0.009

C.3 Experiments under Different Time Discretization

Like other discrete-time NN survival models, DoseSurv supports flexible discretization of survival times into any number of intervals. In our main experiments, we used 30 time intervals, aligning with common practice in the literature [45, 14]. However, DoseSurv can readily operate under different temporal resolutions. Here, we perform additional experiments on data simulated using a finer time discretization of 60 intervals and compare DoseSurv to the continuous-time, CoxPH-inspired DeepSurv model, which may offer greater robustness. Table C.2 presents the corresponding results for data simulated under Scenario S1 ($\eta = 1$). We find that DoseSurv yields metrics comparable to those obtained with 30 intervals in the main experiment, whereas DeepSurv continues to underperform relative to DoseSurv.

Table C.2: Model performance of DoseSurv and DeepSurv under Scenario S1 ($\eta = 1$) with a finer time discretization (60 intervals). For reference, we include performance metrics achieved by DoseSurv in the corresponding main experiment with 30 intervals. Metrics are averaged over 5 runs.

Intervals	Method	$\sqrt{\text{MISE}}$	$\sqrt{\text{DPE}}$	C^{td}	IBS
30	DoseSurv	4.46	7.97	0.75	0.16
60	DoseSurv	5.51	8.00	0.76	0.14
	DeepSurv (CoxPH)	8.44	10.43	0.61	0.20

C.4 Sensitivity Analysis

To assess the robustness of DoseSurv to hyperparameter choices, we conducted a sensitivity analysis under Scenario S1 ($\eta = 1$). We varied the learning rate (lr), batch size, and IPM regularization strength (γ), while keeping all other hyperparameters fixed. Results are averaged over 3 independent runs. The model demonstrates consistent performance across a broad range of settings, indicating relative stability with respect to these key hyperparameters.

Table C.3: Sensitivity of DoseSurv to learning rate, batch size, and IPM regularization under Scenario S1 ($\eta = 1$). Metrics are averaged over 3 runs.

lr	Batch Size	γ	$\sqrt{\text{MISE}}$	$\sqrt{\text{DPE}}$	C^{td}	IBS
0.001	500	0.01	4.58	7.94	0.75	0.16
0.001	500	0.10	4.48	7.88	0.75	0.16
0.001	1000	0.01	4.66	8.09	0.75	0.16
0.001	1000	0.10	4.70	8.07	0.75	0.16
0.005	500	0.01	5.47	8.39	0.73	0.17
0.005	500	0.10	5.73	8.43	0.72	0.17
0.005	1000	0.01	4.81	8.04	0.74	0.16
0.005	1000	0.10	4.95	8.10	0.74	0.16

C.5 Visualizing Dose-Response Curves

Fig. C.2 shows an example dose-response curve predicted by DoseSurv in a single run under scenario S1 ($\eta = 0$) for a selected sample in red. The corresponding ground truth curve is depicted in green. In practice, dose-response curves can be obtained by evaluating the model for a range of dosage values q and using the predicted RMST as the outcome of interest.

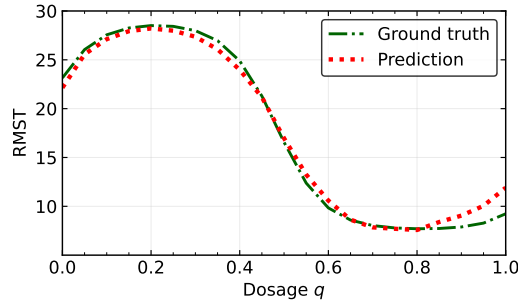


Figure C.2: Example of predicted and ground-truth dose-response curves under scenario S1 ($\eta = 0$).

D Extended Literature Review

D.1 ML Methods for HTE Estimation.

Inferring HTEs of *binary* treatments has been studied extensively in the recent ML literature. Following Curth and van der Schaar [16], methods can typically be categorized into *model-specific* and *model-agnostic* approaches. The latter, also referred to as “meta-learners” [40], can be implemented using any kind of ML method. The most common types of meta-learners (*indirect estimators*) aim to obtain the HTE as the difference between regression estimates of the two potential outcome functions. T-learners [40, 47], fit two (or more) regression models (one for each treatment group) separately, while S-learners [23, 34, 40, 72] fit a single regression model where the treatment variable is added as an extra feature. Meta-learners have been implemented using a wide range of ML models, particularly tree-based models [23, 47, 69, 40, 6] and NNs [34, 72]. On the other hand, *model-specific* approaches rely on unique properties of the underlying ML model. These include approaches which could be described as *hybrids* between S- and T-learners [15]. Such methods are typically implemented as “multi-task learning” models, which learn different treatment responses simultaneously, allowing some information flow between treatments [1, 2, 62, 63, 16]. Beyond architecture, balancing and reweighting strategies are also used to mitigate confounding and covariate shift, including IPM-based representation balancing [62], counterfactual representation learning with balancing weights [5], importance-sampling reweighting [27], and propensity-based regularization [63].

HTE estimation of *continuous-valued* interventions is less explored in the literature. Traditional statistical approaches typically rely on the generalized propensity score [30]. The dose-response-networks proposed by Schwab et al. [60] extend the multi-task learning NN architecture for binary treatments, first presented by Shalit et al. [62], by subdividing the treatment-specific layers, into multiple network heads, each assigned to a different dosage interval. This facilitates the regression of the dose-response function across different dosage regions. A drawback of this architecture is that bin-specific heads produce separate outputs for different dosage intervals, breaking the continuity in the dose-response curves. Inspired by [28], Nie et al. [51] aim to overcome this limitation by employing a varying-coefficient network (VCNet) in the prediction head. Here, the network weights are modeled as continuous functions of the dosage. Schweisthal et al. [61] extended this approach to learn the effect of combinations of multiple dosage-dependent treatments. A different approach by Bica et al. [9] extends the GAN framework proposed by Yoon et al. [70], and introduced a hierarchical discriminator specifically tailored for continuous-valued treatment settings. However, leveraging adversarial learning for counterfactual survival analysis is challenging due to censoring and time structure, which complicate defining faithful discrimination targets and stable training objectives. Existing balancing methods for continuous-valued treatments assume clustered data [11], which limits its applicability in settings where natural clusters are absent. Moreover, Bellot et al. [8] provide a generalization bound, but their method relies on estimating complex integrals over representation and treatment spaces, which makes it likely to be sensitive to function and kernel choices.

D.2 Survival Analysis and HTE Inference with TTE Data.

ML methods for survival analysis have gained significant attention in recent years. Notable approaches include adaptations of the Cox-proportional hazards model [13], where traditional linear predictors are replaced with feedforward NNs [20, 37]. In addition, TTE has been modeled via accelerated failure time (AFT) regression [35] and deep generative models [55]. Random survival forests have also emerged as a flexible approach to predict cumulative hazard functions in a non-parametric tree-ensemble manner [32]. Deep learning-based classifiers [10, 24, 56, 41] have gained popularity for this purpose. Alternatively, Lee et al. [45] parameterize the likelihood in discrete-time settings using the probability mass function. Finally, survival clustering methods have been proposed to identify groups of samples based on their survival patterns [7, 33].

Despite advances in survival analysis, adapting these models for HTE estimation remains understudied. Early investigations focus primarily on binary treatment settings, leveraging tree-based methods such as causal trees [71], Bayesian additive regression trees (BART) [29], and random forests [67] to measure the effect of binary treatments on expected survival time. More recent approaches, such as that by Chapfuwa et al. [12], explore representation learning and generative modeling to estimate event times for different treatments. A seminal work by Curth et al. [14] adopts a multi-task NN architecture with time- and treatment-specific network heads to predict discrete conditional hazard

functions, incorporating balanced representations to address covariate shifts. Schrod et al. [58] simplify this approach by combining a dichotomous multi-task architecture with a Cox-proportional hazard loss [37]. However, such works focus on binary interventions. In comparison, our DoseSurv model is more versatile and can handle single- and multi-arm treatment regimens with mutually exclusive continuous-valued (dosage-dependent) and non-dosage treatment options.

D.3 Comparison of DoseSurv with Related Work

Table D.1 compares key ML methods for HTE estimation, specifically targeting continuous-valued treatments or TTE outcomes. Unlike related work, DoseSurv uniquely integrates TTE analysis with complex treatment settings that include continuous-valued treatments, non-dosage treatments, and mixed treatment scenarios. Due to the lack of HTE estimation models for such settings, our baseline comparisons rely on standard survival models, implemented as model-agnostic T- and S-learners. Although not explicitly studied here, DoseSurv is generally also applicable to treatment scenarios with more than two treatment options.

Table D.1: Overview of selected key neural methods for HTE estimation under continuous-valued treatments or TTE outcomes and their applicability across different settings.

Model	Continuous-valued (Dosage) Treatments	Mixed Dosage and Non-dosage Options	More than Two Treatment Options	TTE Outcomes
DRNet [60]	✓	✗	✓	✗
SCIGAN [9]	✓	(✓)	✓	✗
VCNet [51]	✓	✗	✗	✗
SurvITE [14]	✗	✗	✗	✓
BITES [58]	✗	✗	✗	✓
DoseSurv (ours)	✓	✓	✓	✓

E Implementation

E.1 DoseSurv

The source code of DoseSurv will be made available at: <https://github.com/mgoegl/DoseSurv>.

Experiments were performed on a single GPU machine (specifications: CPU – Intel Xeon W5-2445, GPU – NVIDIA RTX A4000, and RAM – 64GB DDR5, OS – Linux). DoseSurv is implemented in PyTorch. The training time of DoseSurv was under 30 seconds per run for Scenario S1 ($\eta = 0$) without IPM regularization. Per-epoch training time generally ranges from under a second to a few seconds, depending on whether IPM regularization is applied. Overall, computation time increases with the number of IPM bins.

Across all main experiments, we implemented DoseSurv with 1 hidden layer in the representation network and 3 hidden layers in each treatment-specific head of the RBF hazard estimator, each layer comprising 100 nodes. For each layer in the RBF hazard estimator, we use 5 Gaussian RBFs with centers, initialized at $\{0, 0.25, 0.5, 0.75, 1\}$. The shape parameters $\sigma_{l,k}^{(a)}$ are initialized at 0.7. For each experiment, we chose γ from $\{0.1, 0.01\}$ based on the lowest INBLL achieved on the validation data. For experiments on synthetic data, we optimized γ under $\eta = 2$. Network parameters were optimized using Adam optimizer [38] with learning rate of 0.001. We use ReLU activation functions, dropout probability of 0.1, and a batch size of 500. Moreover, we implement DoseSurv with batch normalization layers, and employ early stopping after 30 epochs without model improvement on the validation data. We adopt the implementation of the finite-sample approximation of the Wasserstein distance from [14], which in turn follows the implementation in [5, 17], and improve computational efficiency by parallelizing its computation across treatments a , dosage bins q , and time bins t . We use the same parameters, i.e., an entropic regularization strength of $\lambda_{\text{wass}} = 10$ and 10 Sinkhorn iterations. Moreover, we choose $\mathfrak{Q} = 3$ and $\mathfrak{T} = 5$.

Deviating from the primary experiments, we configure DoseSurv for the additional experiment described in C.2 with a representation network consisting of 3 layers of 200 nodes each and a

representation size of 50. The treatment-specific hazard estimators are designed with 2 layers of 100 nodes each, without applying batch normalization. Moreover, we choose $\mathfrak{L} = 1$ and $\mathfrak{T} = 10$.

E.2 Baseline Models

We benchmarked DoseSurv against publicly available implementations of four state-of-the-art machine and deep learning models for survival analysis, **DeepSurv**⁸, **DeepHit**⁸, **RSF**⁹ and **NSC**¹⁰. We adopt these implementations as both T- and S-learners. First, we standardize covariates and treatment-specific dosages. For the standard (S-learner) baselines, we use slot-based (continuous one-hot) encoding of treatment and dosage information. We add the dosage information for each treatment option as additional continuous covariates. In scenarios with more than one treatment option, we set the dosage values for unobserved treatments to -10 , which does not normally fall within the standardized dosage range, ensuring that the models do not mistakenly interpret this value as a valid dosage. For categorical (non-dosage) treatments, we set the dosage parameter for the observed treatment to 1. In addition, for scenarios S2 and S3 with more than one treatment group, we implement T-learner versions of the baseline models: **DeepSurv-T**, **DeepHit-T**, **RSF-T**, and **NSC-T**. Here, we split the data into the respective treatment groups, and train separate versions of the model on each treatment group individually, including the corresponding dosage as an additional covariate.

We optimize hyperparameters based on a grid search. We choose the optimal hyperparameter combination based on the minimum INBLL, which comprises both calibration and discriminative performance. For DeepSurv we optimize the number of layers and nodes, choosing from $\{[100], [100, 100], [100, 100, 100], [200, 200]\}$. Additionally, for DeepHit, we optimize the parameters $\alpha_{\text{DeepHit}} \in \{0.2, 0.5\}$ and $\sigma_{\text{DeepHit}} \in \{0.1, 1\}$, which control the contribution and properties of the ranking loss. For NSC, we choose network sizes of the mixture weights and survival networks from $\{[100, 100, 100], [200, 200]\}$. Furthermore, we choose the number of components for the mixture from $\{4, 6\}$, and the size of the latent cluster representation from $\{10, 100\}$. For RSF, we optimize the number of trees $\{100, 300\}$ as well as the minimum numbers of samples required to split an internal node ($\{6, 12\}$) or be at a leaf node ($\{3, 6\}$). For all network-based baseline models, as for DoseSurv, we use a dropout rate of 0.1, a batch size of 500, batch normalization, and early stopping after 30 epochs.

For experiments on the Twins and breast cancer datasets under binary treatments, we additionally compare against adapted public implementations of the HTE models **BITES**¹¹ and **SurvITE**¹² to ensure compatibility with our environment. For SurvITE, we choose the standard network architecture (representation network: $[100, 100, 100]$; hazard estimators: $[100, 100]$). For BITES, we choose the standard representation network of size $[7, 5]$, and optimize the treatment-specific heads from $\{[5, 3], [3]\}$. As for DoseSurv, we choose the IPM regularization strength for both models from $\{0.01, 0.1\}$ and BITES’s blur parameter from $\{0.05, 0.1\}$.

F Remarks on HTE Estimation and Assumptions

In HTE inference, we generally aim to estimate

$$\begin{aligned}\nu(\mathbf{x}) &= \mathbb{E}[Y_1 - Y_0 | \mathbf{X} = \mathbf{x}] \\ &= \mathbb{E}[Y_1 | \mathbf{X} = \mathbf{x}] - \mathbb{E}[Y_0 | \mathbf{X} = \mathbf{x}]\end{aligned}$$

where Y_0 and Y_1 are two continuous potential outcomes under treatments 0 and 1, and $\mathbf{X} = \mathbf{x}$ are the patient covariates. The potential outcomes can also be expressed using the do-operator [53] indicating an intervention to be made:

$$\nu(\mathbf{x}) = \mathbb{E}[Y | \text{do}(A=1), \mathbf{X} = \mathbf{x}] - \mathbb{E}[Y | \text{do}(A=0), \mathbf{X} = \mathbf{x}].$$

⁸<https://github.com/havakv/pycox>

⁹scikit-survival [54]

¹⁰<https://github.com/Jeanselme/NeuralSurvivalClustering>

¹¹<https://github.com/sschrod/BITES>

¹²<https://github.com/chl8856/survITE>

For HTE estimation between two dosage-dependent interventions (a, q) and $(\tilde{a}, \tilde{q})$ and the time-specific survival probabilities as outcome of interest, this translates into

$$\begin{aligned}\nu_{\text{surv}}(t|\mathbf{x}) &= \mathbb{P}(T_{a,q} > t | \mathbf{X} = \mathbf{x}, \text{do}(C \geq t)) - \\ &\quad \mathbb{P}(T_{\tilde{a},\tilde{q}} > t | \mathbf{X} = \mathbf{x}, \text{do}(C \geq t)) \\ &= S_{a,q}(t|\mathbf{x}) - S_{\tilde{a},\tilde{q}}(t|\mathbf{x}).\end{aligned}$$

However, since ground-truth survival functions are unknown in real-world settings, we focus on estimating the HTE in terms of the RMST (up until the time horizon, t_m) given by

$$\begin{aligned}\nu_{\text{rmst}}(\mathbf{x}) &= \mathbb{E}[\min(T_{a,q}, t_m) | \mathbf{X} = \mathbf{x}] - \mathbb{E}[\min(T_{\tilde{a},\tilde{q}}, t_m) | \mathbf{X} = \mathbf{x}] \\ &= t_{a,q}^{\text{rmst}}(\mathbf{x}) - t_{\tilde{a},\tilde{q}}^{\text{rmst}}(\mathbf{x}) \\ &= \sum_{t=0}^{t_m-1} S_{a,q}(t|\mathbf{x}) - \sum_{t=0}^{t_m-1} S_{\tilde{a},\tilde{q}}(t|\mathbf{x}) \\ &= \sum_{t=0}^{t_m-1} \nu_{\text{surv}}(t|\mathbf{x}).\end{aligned}$$

Since $S_{a,q}(t|\mathbf{x}) = \prod_{\tau \leq t} (1 - \lambda_{a,q}(\tau|\mathbf{x}))$, we therefore need to compute $\lambda_{a,q}(\tau|\mathbf{x})$ in order to find unbiased estimates for both $\nu_{\text{surv}}(t|\mathbf{x})$ and $\nu_{\text{rmst}}(\mathbf{x})$. We can rewrite

$$\begin{aligned}\lambda_{a,q}(\tau|\mathbf{x}) &= \\ &= \mathbb{P}(T = \tau | T \geq \tau, \mathbf{X} = \mathbf{x}, \text{do}(A = a, Q = q, C \geq \tau)) \\ &= \mathbb{P}(T_{a,q} = \tau | T_{a,q} \geq \tau, \mathbf{X} = \mathbf{x}, \text{do}(C \geq \tau)).\end{aligned}$$

Using Assumptions 1, 3–4, outlined in Section 3.2, we obtain [14]

$$\begin{aligned}\lambda_{a,q}(\tau|\mathbf{x}) &= \\ &\stackrel{(1)}{=} \mathbb{P}(T_{a,q} = \tau | T_{a,q} \geq \tau, \mathbf{X} = \mathbf{x}, A = a, Q = q, \text{do}(C \geq \tau)) \\ &\stackrel{(4)}{=} \mathbb{P}(T_{a,q} = \tau | T_{a,q} \geq \tau, \mathbf{X} = \mathbf{x}, A = a, Q = q, C \geq \tau) \\ &\stackrel{(3)}{=} \mathbb{P}(T = \tau | T \geq \tau, \mathbf{X} = \mathbf{x}, A = a, Q = q, C \geq \tau) \\ &= \mathbb{P}(\tilde{T} = \tau, \Delta = 1 | \mathbf{X} = \mathbf{x}, A = a, Q = q, \tilde{T} \geq \tau) \\ &= \lambda(\tau|\mathbf{x}, a, q),\end{aligned}$$

which can be estimated non-parametrically under the additional assumptions that treatment and dosage assignment (Assumption 2) and the survival and censoring processes (Assumption 5) satisfy positivity (i.e., are non-deterministic).

G Causal Diagram

The directed acyclic graph (DAG) [53] for our setting is shown in Fig. G.1 and follows prior work on HTE estimation with TTE data [14, 65], extended here to include the dosage Q . We define $\mathcal{N}(t) = \mathbb{1}(\tilde{T} \leq t, \Delta = 1)$ and $\mathcal{N}_C(t) = \mathbb{1}(\tilde{T} \leq t, \Delta = 0)$ as counting processes indicating, over time, whether an event (death) or, respectively, censoring has occurred. By definition, $\mathcal{N}(t)$ and $\mathcal{N}_C(t)$ are 0 for $t = 0$ and switch to 1 for $t \geq \tilde{T}$, i.e., when either the event or censoring occurs at time $\tilde{T}$.

$\mathbf{X}$ consists of (possibly overlapping) subgroups $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3$, and $\mathbf{X}_4$ of covariates that influence (1) the event process, (2) treatment assignment, (3) dosage assignment, and (4) the censoring process. These subgroups and their causal influences are associated with covariate shifts arising from (1) event-induced bias, (2) treatment-selection bias, (3) dosage-selection bias, and (4) censoring bias, respectively [14].

The causal diagram describes how $\mathcal{N}$ and $\mathcal{N}_C$ evolve over time under the influence of covariates $\mathbf{X}$, treatment A , and dosage Q , and is in line with the Assumptions formulated in Section 3.2. The

temporal dependencies shown ($\mathcal{N}(t-1) \rightarrow \mathcal{N}(t)$ and $\mathcal{N}_C(t-1) \rightarrow \mathcal{N}(t), \mathcal{N}_C(t)$) reflect the fact that these counting processes are sequential—an individual’s event status at time t depends on their status at the previous time step $t-1$. In other words: once $\mathcal{N}(t-1) = 1$ (event occurred by time $t-1$), then $\mathcal{N}(t) = 1$; similarly, once $\mathcal{N}_C(t-1) = 1$ (censored by time $t-1$), then $\mathcal{N}_C(t) = 1$ and $\mathcal{N}(t) = 0$ (no event can occur after censoring). Furthermore, $\mathcal{N}(t) \rightarrow \mathcal{N}_C(t)$ captures the logical constraint that an event precludes censoring—if $\mathcal{N}(t) = 1$, then $\mathcal{N}_C(t) = 0$.

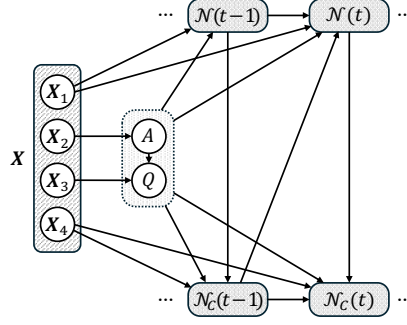


Figure G.1: Assumed DAG for our discrete-time survival analysis setting with continuous-valued treatment, extending prior formulations [14, 65].

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:

- We address the problem of HTE estimation for single- and multi-arm treatment regimens with continuous-valued (dosage-dependent) and non-dosage treatment options, as well as TTE outcomes—a practically important but largely underexplored area. (See Section [3.](#))
- We propose a novel varying-coefficient network with layer-specific RBFs, specifically designed for HTE estimation in the above setting. (See Section [4.](#))
- We introduce a computationally efficient IPM regularization approach to address biases in HTE estimation from observational TTE data with continuous-valued treatments. (See Section [4.](#))
- We validate our model across diverse scenarios and datasets, including a reinterpretation of the classical Twins dataset, serving as a real-world benchmark for continuous-valued treatments with TTE outcomes. (See Section [5.](#))

Guidelines:

- The answer NA 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 NA 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: We discuss limitations of the work in Section [6.](#)

Guidelines:

- The answer NA 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: All assumptions are explicitly stated in Section 3.2. Further mathematical derivations related to estimating the HTE of continuous-valued interventions from TTE data and the corresponding assumptions, are provided in Appendix F.

Guidelines:

- The answer NA 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: We present the experimental settings in Section 5 and provide implementation details and hyperparameters for our model and baselines in Appendix E. The data-generation processes are described in Appendix A. Reviewers can find the source code in the supplemental material.

Guidelines:

- The answer NA 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: Reviewers can find the source code in the supplemental material. All raw real-world datasets are publicly available. Additionally, the code will be made available on GitHub later.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.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://nips.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, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We provide these details in Appendix E.

Guidelines:

- The answer NA 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: All main experiments are repeated across multiple runs. We report 95% confidence intervals to capture variability—except in Fig. 2 and Fig. 3, where they are omitted for visual clarity.

Guidelines:

- The answer NA 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: We provide the details of computer resources used in Appendix E.

Guidelines:

- The answer NA 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: We conform to the NeurIPS Code of Ethics.

Guidelines:

- The answer NA 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: [\[Yes\]](#)

Justification: We discuss the potential societal impact of our work in Section 6.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA 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., pretrained language models, image generators, or scraped datasets)?

Answer: [\[NA\]](#) .

Justification: No such risks for our paper.

Guidelines:

- The answer NA 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: [Yes]

Justification: We cite the original papers that are used for our code implementation and discuss details in Appendix E.

Guidelines:

- The answer NA 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: We provide implementation details in Appendix E and describe the generation process of the (synthetic) datasets we produce in Appendix A.

Guidelines:

- The answer NA 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: [NA] .

Justification: This paper does not involve crowdsourcing or research with human subjects.

Guidelines:

- The answer NA 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: [NA] .

Justification: This research does not involve human subjects.

Guidelines:

- The answer NA 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: [NA] .

Justification: The usage of LLMs did not form an original, or non-standard component of the core method.

Guidelines:

- The answer NA 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 (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.