ENABLING COUNTERFACTUAL SURVIVAL ANALYSIS WITH BALANCED REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Balanced representation learning methods have been applied successfully to counterfactual inference from observational data. However, approaches that account for survival outcomes are relatively limited. Survival data are frequently encountered across diverse medical applications, *i.e.*, drug development, risk profiling, and clinical trials, and such data are also relevant in fields like manufacturing (for equipment monitoring). When the outcome of interest is time-to-event, special precautions for handling censored events need to be taken, as ignoring censored outcomes may lead to biased estimates. We propose a theoretically grounded unified framework for counterfactual inference applicable to survival outcomes. Further, we formulate a nonparametric hazard ratio metric for evaluating average and individualized treatment effects. Experimental results on real-world and semi-synthetic datasets, the latter which we introduce, demonstrate that the proposed approach significantly outperforms competitive alternatives in both survival-outcome predictions and treatment-effect estimation.

1 INTRODUCTION

Survival analysis or time-to-event studies focus on modeling the time of a future event, such as death or failure, and investigate its relationship with covariates or predictors of interest. Specifically, we may be interested in the *causal effect* of a given intervention or treatment on survival time. A typical question may be: will a given therapy increase the chances of survival of an individual or population? Such causal inquiries on survival outcomes are common in the fields of epidemiology and medicine (Robins, 1986; Hammer et al., 1996; Yusuf et al., 2016). As an important current example, the COVID-19 pandemic is creating a demand for methodological development to address such questions, specifically, when evaluating the effectiveness of a potential vaccine or therapeutic outside randomized controlled trial settings.

Traditional causal survival analysis is typically carried out in the context of a randomized controlled trial (RCT), where the treatment assignment is controlled by researchers. Though they are the gold standard for causal inference, RCTs are usually long-term engagements, expensive and limited in sample size. Alternatively, the availability of *observational* data with comprehensive information about patients, such as electronic health records (EHRs), constitutes a more accessible but also more challenging source for estimating causal effects (Häyrinen et al., 2008; Jha et al., 2009). Such observational data may be used to augment and verify an RCT, after a particular treatment is approved and in use (Gombar et al., 2019; Frankovich et al., 2011; Longhurst et al., 2014). Moreover, the wealth of information from observational data also allows for the estimation of the individualized treatment effect (ITE), namely, the causal effect of an intervention at the individual level. In this work, we develop a novel framework for *counterfactual time-to-event prediction* to estimate the ITE for survival or time-to-event outcomes from observational data.

Estimating the causal effect for survival outcomes in observational data manifests two principal challenges. First, the treatment assignment mechanism is not known *a priori*. Therefore, there may be variables, known as *confounders*, affecting both the treatment and survival time, which lead to selection bias (Bareinboim & Pearl, 2012), *i.e.*, that the distributions across treatment groups are not the same. For instance, patients who are severely ill are likely to receive more aggressive therapy, however, their health status may *also* inevitably influence survival. Traditional survival analysis neglects such bias, leading to incorrect causal estimation. Second, the exact time-to-event is not

always observed, *i.e.*, sometimes we only know that an event has *not* occurred up to a certain point in time. This is known as the *censoring* problem. Moreover, censoring might be informative depending on the characteristics of the individuals and their treatment assignments, thus proper adjustment is required for accurate causal estimation (Cole & Hernán, 2004; Díaz, 2019).

Traditional causal survival-analysis approaches typically model the effect of the treatment or covariates (not time or survival) in a parametric manner. Two commonly used models are the Cox proportional hazards (CoxPH) model (Cox, 1972) and the accelerated failure time (AFT) model (Wei, 1992), which presume a linear relationship between the covariates and survival probability. Further, proper weighting for each individual has been employed to remove confounding bias from these models (Austin, 2007; 2014; Hernán et al., 2005). For instance, probability weighting schemes that account for both selection bias and covariate dependent censoring have been considered for adjusted survival curves (Cole & Hernán, 2004; Díaz, 2019). Moreover, such probability weighting schemes have been applied to causal survival-analysis under time-varying treatment and confounding (Robins, 1986; Hernán et al., 2000). See van der Laan & Robins (2003); Tsiatis (2007); Van der Laan & Rose (2011); Hernán & Robins (2020) for an overview. Such linear specification makes these models interpretable but compromises their flexibility, and makes it difficult to adapt them for high-dimensional data or to capture complex interactions among covariates. Importantly, these methods lack a counterfactual prediction mechanism, which is key for ITE estimation.

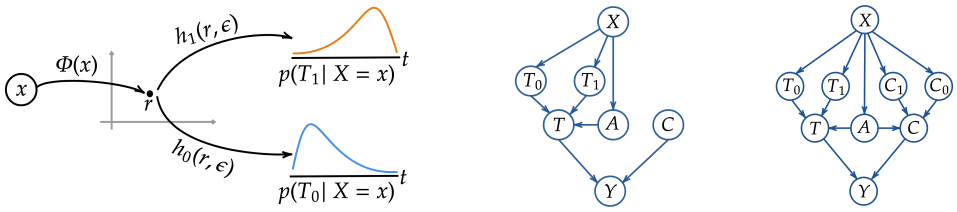
Fortunately, recent advances in machine learning, such as representation learning or generative modeling, have enabled causal inference methods to handle high-dimensional data and to characterize complex interactions effectively. For instance, there has been recent interest in tree-based (Chipman et al., 2010; Wager & Athey, 2018) and neural-network-based (Shalit et al., 2017; Zhang et al., 2020) approaches. For pre-specified time-horizons, the nonparametric Random Survival Forest (RSF) (Ishwaran et al., 2008) and Bayesian Additive regression trees (BART) (Chipman et al., 2010) have been extended to causal survival analysis. RSF has been applied to causal survival forests with weighted bootstrap inference (Shen et al., 2018; Cui et al., 2020) while a BART is extended to account for survival outcomes in Surv-BART (Sparapani et al., 2016) and AFT-BART (Henderson et al., 2020). Alternatively, when estimating the ITE, neural-network-based methods propose to regularize the transformed covariates or representations for an individual to have balanced distributions across treatment groups, thus removing the confounding bias and improving ITE prediction. However, most approaches employing *representation learning* techniques for counterfactual inference deal with continuous or binary outcomes, instead of time-to-event outcomes with censoring. Hence, a principled generalization to the context of *counterfactual survival analysis* is needed.

In this work we leverage balanced (latent) representation learning to estimate ITE via counterfactual prediction of survival outcomes in observational studies. We develop a framework to predict event times from a low-dimensional transformation of the original covariate space. To address the specific challenges associated with counterfactual survival analysis, we make the following contributions:

- We develop an optimization objective incorporating adjustments for informative censoring, as well as a balanced regularization term bounding the generalization error for ITE prediction. For the latter, we repurpose a recently proposed bound (Shalit et al., 2017) for our time-to-event scenario.
- We propose a generative model for event times to relax restrictive survival linear and parametric assumptions, thus allowing for more flexible modeling. Our approach can also provide nonparametric uncertainty quantification for ITE predictions.
- We provide survival-specific evaluation metrics, including a new *nonparametric hazard ratio* estimator, and discuss how to perform model selection for survival outcomes. The proposed model demonstrates superior performance relative to the commonly used baselines in real-world and semi-synthetic datasets.
- We introduce a survival-specific semi-synthetic dataset and demonstrate an approach for leveraging prior randomized experiments in longitudinal studies for model validation.

2 Problem Formulation

We first introduce the basic setup for performing causal survival analysis in observational studies. Suppose we have N units, with N_1 units being *treated* and N_0 in the *control* group ($N = N_1 + N_0$). For each unit (individual), we have covariates X, which can be heterogeneous, *e.g.*, a mixture of categorical and continuous covariates which, in the context of medicine, may include labs, vitals,



(a) Counterfactual inference (b) Non-informative (c) Informative

Figure 1: (a) Illustration of the proposed counterfactual survival analysis (CSA). Covariates $X = x$ are mapped into latent representation r via deterministic mapping $r = \Phi(x)$. The potential outcomes are sampled from $t_a \sim p(T_A|X = x)$ for $A = a$ via stochastic mapping $h_A(r, \tilde{\epsilon})$, where randomness is induced with a flow-based transformation, $\tilde{\epsilon}$, of a simple distribution $p(\epsilon)$, *i.e.*, uniform or Gaussian. (b) and (c) show the proposed causal graphs for non-informative and informative censoring, respectively.

procedure codes, *etc*. We also have a *treatment* indicator A, where $A = 0$ for the controls and $A = 1$ for the treated, as well as the outcome (event) of interest T. Under the potential-outcomes framework (Rubin, 2005), let T_0 and T_1 be the potential event times for a given subject under control and treatment, respectively. In practice we only observe one realization of the potential outcomes, *i.e.*, the *factual* outcome $T = T_A$, while the *counterfactual* outcome T_{1-A} is unobserved.

In survival analysis, the problem becomes more difficult because we do *not* always observe the exact event time for each individual, but rather the time up to which we are certain that the event has not occurred; specifically, we have a (right) censoring problem, most likely due to the loss of follow-up. We denote the censoring time as C and censoring indicator as $\delta \in \{0, 1\}$. The actual *observed time* is $Y = \min(T_A, C)$, *i.e.*, the outcome is observed (non-censored) if $T_A < C$ and $\delta = 1$.

In this work, we are interested in the expected difference between the T_1 and T_0 conditioned on X for a given unit (individual), which is commonly known as the *individualized treatment effect* (ITE). Specifically, we wish to perform inference on the conditional distributions of T_1 and T_0, *i.e.*, $p(T_1|X)$ and $p(T_0|X)$, respectively, as shown in Figure 1a. In practice, we observe N realizations of (Y, δ, X, A) for observed time, censoring indicator, covariates and treatment indicator, respectively; hence, from an observational study the dataset takes the form $\mathcal{D} = \{(y_i, \delta_i, x_i, a_i)\}_{i=1}^N$. Below, we discuss several common choices of estimands in survival analysis.

Estimands of Interest We begin by considering survival analysis in the *absence* of an intervening treatment choice, A. Let $F(t|x) \triangleq P(T \leq t|X = x)$ be the cumulative distribution function of the event (failure) time, t, given a realization of the covariates, x. Survival analysis is primarily concerned with characterization of the *survival function* conditioned on covariates $S(t|x) \triangleq 1 - F(t|x)$, and the *hazard function* or *risk score*, $\lambda(t|x)$, defined below. $S(t|x)$ is a monotonically decreasing function indicating the probability of survival up to time t. The hazard function measures the instantaneous probability of the event occurring between $\{t, t + \Delta t\}$ given $T > t$ and $\Delta t \to 0$. From standard definitions (Kleinbaum & Klein, 2010), the relationship between cumulative and hazard function is $f(t|x) \triangleq P(T = t|X = x) = \lambda(t|x)S(t|x)$, *i.e.*, the conditional *event time density function*.

Given the binary treatment A, we are interested in its impact on the survival time. For ITE estimation, we are also interested in the difference between the two potential outcomes T_1, T_0. Let $S_A(t|x)$ and $\lambda_A(t|x)$ denote the survival and hazard functions for the potential outcomes T_A, *i.e.*, T_1 and T_0. Several common estimands of interest include (Zhao et al., 2012; Trinquart et al., 2016): *difference in expected lifetime*: $\text{ITE}(t, x) = \int_0^{t_{\max}} \{S_1(t|x) - S_0(t|x)\} \mathrm{d}t = \mathbb{E}\{T_1 - T_0|X = x\}$, *difference in survival function*: $\text{ITE}(t, x) = S_1(t|x) - S_0(t|x)$, and *hazard ratio*: $\text{ITE}(t, x) = \lambda_1(t|x)/\lambda_0(t|x)$. The inference difficulties associated with the above estimands from observational data are two-fold. First, there are confounders affecting both the treatment assignment and outcomes, which stem from selection bias, *i.e.*, the treatment and control covariate distributions are not necessarily the same. Also, we do not have direct knowledge of the conditional treatment assignment mechanism, *i.e.*, $P(A = a|X = x)$, also known as the *propensity score*. Let $\perp\!\!\!\perp$ denote statistical independence. For estimands to be identifiable from observational data, we make two assumptions: (i) $\{T_1, T_0\} \perp\!\!\!\perp A|X$, *i.e.*, no unobserved confounders or *ignorability*, and (ii) *overlap* in the covariate support $0 < P(A = 1|X = x) < 1$ *almost surely if* $p(X = x) > 0$. Second, the censoring mechanism is also unknown and may lead to bias without proper adjustment. We consider two censoring mechanisms in our

work, (i) conditionally independent or *informative censoring*: $T \perp\!\!\!\perp C|X, A$, and (ii) random or *non-informative censoring*: $T \perp\!\!\!\perp C$. Note that for informative censoring, we also have to consider potential censoring times C_1 and C_0 and their conditionals $p(C_1|X)$ and $p(C_0|X)$, respectively. Figure 1 shows causal graphs illustrating these modeling assumptions.

3 MODELING

To overcome the above challenges and adjust for observational biases, we propose a unified framework for *counterfactual survival analysis* (CSA). Specifically, we repurpose the counterfactual bound in Shalit et al. (2017) for our time-to-event scenario and introduce a nonparametric approach for stochastic survival outcome predictions based on a planar normalizing flow (Rezende & Mohamed, 2015). Below we formulate a theoretically grounded and unified approach for estimating (i) the encoder function $r = \Phi(x)$, which deterministically maps covariates x to their corresponding latent representation $r \in \mathbb{R}^d$, and (ii) two stochastic time-to-event generative functions, $h_A(\cdot)$, to implicitly draw samples from both potential outcome conditionals $t_a \sim p_{h,\Phi}(T_A|X = x)$, for $A = \{1, 0\}$, and where t_a indicates the sample from $p_{h,\Phi}(T_A|X = x)$ is for $A = a$. Further, we formulate a general extension that accounts for informative censoring by introducing two stochastic censoring generative functions, $\nu_A(\cdot)$, to draw samples for potential censoring times $c_a \sim p_{\nu,\Phi}(C_A|X = x)$. The model-specifying functions, $\{h_A(\cdot), \nu_A(\cdot), \Phi(\cdot)\}$, are parameterized via neural networks. See the Supplementary Material (SM) for details. Figure 1a summarizes our modeling approach.

Accounting for selection bias We wish to estimate the treatment effect from intervention $do(A = a)$ (Pearl & Bareinboim, 2014), where the potential outcomes, event times, are sampled from distributions parameterized by functions $\{h_A(\cdot), \Phi(\cdot)\}$, *i.e.*,

$$t \sim p_{h,\Phi}(T|X = x, do(A = a)) \tag{1}$$

$$t \sim p_{h,\Phi}(T|X = x, A = a) \tag{2}$$

$$t_a \sim p_{h,\Phi}(T_A|X = x), \tag{3}$$

where (2) follows from (1) because *identifiability* holds, provided that X is a sufficient set from A into T (Pearl & Bareinboim, 2014), consistent with the causal graphs in Figure 1b and 1c. We obtain (3) from (2) via the *strong ignorability* assumption, *i.e.*, $\{T_0, T_1\} \perp\!\!\!\perp A|X$ and $0 < P(A = a|X = x) < 1$, and the *consistency* assumption, *i.e.*, $T = T_A|A = a$. A similar argument can be made for informative censoring based on Figure 1c, so we can also write $c_a \sim p_{\nu,\Phi}(C_A|X = x)$. Given (3), model functions $\{h_A(\cdot), \Phi(\cdot)\}$ and $\nu_A(\cdot)$ for informative censoring can be learned by leveraging standard statistical optimization approaches, that minimize a loss hypothesis $\mathcal{L}$ given samples from the empirical distribution $(y, \delta, x, a) \sim p(Y, \delta, X, A)$, *i.e.*, from dataset $\mathcal{D}$. Specifically, we write $\mathcal{L} = \mathbb{E}_{(y,\delta,x,a)\sim p(Y,\delta,X,A)}\left[\ell_{h,\Phi}(t_a, y, \delta)\right]$, where $\ell_{h,\Phi}(t_a, y, \delta)$ is a loss function that measures the agreement of $t_a \sim p_{h,\Phi}(T_A|X = x)$ (and $c_a \sim p_{\nu,\Phi}(C_A|X = x)$ for informative censoring) with ground truth $\{y, \delta\}$, the observed time and censoring indicator, respectively.

For some parametric formulations of event time distribution $p_{h,\Phi}(T_A|X = x)$, *e.g.*, exponential, Weibull, log-Normal, *etc.*, and provided the censoring mechanism is non-informative, $-\ell_{h,\Phi}(t_a, y, \delta)$ is the closed form log likelihood. Specifically, $-\ell_{h,\Phi}(t_a, y, \delta) \triangleq \log p_{h,\Phi}(T_a|X = x) = \delta \cdot \log f_{h,\Phi}(t_a|x) + (1 - \delta) \cdot \log S_{h,\Phi}(t_a|x)$, which implies that the conditional event time density and survival functions can be calculated in closed form from transformations $\{h_A(\cdot), \Phi(\cdot)\}$ of x. See the SM for parametric examples of $\mathcal{L}$ accounting for informative censoring.

We further define the expected loss for a given realization of covariates x and treatment assignment a over observed times y (censored and non-censored), and the censoring indicator δ as $\zeta_{h,\Phi}(x, a) \triangleq \mathbb{E}_{(y,\delta,x)\sim p(Y,\delta|X)}\ell_{h,\Phi}(t_a, y, \delta)$ as in Shalit et al. (2017). For a given subject with covariates x and treatment assignment a, we wish to minimize both the factual and counterfactual losses, $\mathcal{L}_{\text{F}}$ and $\mathcal{L}_{\text{CF}}$, respectively, by decomposing $\mathcal{L} = \mathcal{L}_{\text{F}} + \mathcal{L}_{\text{CF}}$ as follows

$$\mathcal{L}_{\text{F}} = \mathbb{E}_{(x,a)\sim p(A,X)}\zeta_{h,\Phi}(x, a)\,, \quad \mathcal{L}_{\text{CF}} = \mathbb{E}_{(x,a)\sim p(1-A,X)}\zeta_{h,\Phi}(x, a)\,. \tag{4}$$

Let $u \triangleq P(A = 1)$ denote the marginal probability of treatment assignment. We can readily decompose the losses in (4) according to treatment assignments. The decomposed factual $\mathcal{L}_{\text{F}} = u \cdot \mathcal{L}_{\text{F}}^{A=1} + (1 - u) \cdot \mathcal{L}_{\text{F}}^{A=0}$, and similarly, the decomposed counterfactual $\mathcal{L}_{\text{CF}} = (1 - u) \cdot \mathcal{L}_{\text{CF}}^{A=1} + u \cdot \mathcal{L}_{\text{CF}}^{A=0}$. In practice, only *factual* outcomes are observed, hence, for a non-randomized non-controlled experiment, we cannot obtain an unbiased estimate of $\mathcal{L}_{\text{CF}}$ from data due to selection bias (or confounding). Therefore, we bound $\mathcal{L}_{\text{CF}}$ and $\mathcal{L}$ below following Shalit et al. (2017).

Corollary 1 *Assume $\Phi(\cdot)$ is an invertible map, and $\alpha^{-1}\zeta_{h,\Phi}(x,a) \in G$, where G is a family of functions, $p_\Phi^{A=a} \triangleq p_\Phi(R|A=a)$ is the latent distribution for group $A=a$, and $\alpha > 0$ is a constant. Then, we have:*

$$\mathcal{L}_{\mathrm{CF}} \leq (1-u)\cdot \mathcal{L}_{\mathrm{F}}^{A=1} + u \cdot \mathcal{L}_{\mathrm{F}}^{A=0} + \alpha \cdot \mathrm{IPM}_G(p_\Phi^{A=1}, p_\Phi^{A=0})$$
$$\mathcal{L} \leq \mathcal{L}_{\mathrm{F}}^{A=1} + \mathcal{L}_{\mathrm{F}}^{A=0} + \alpha \cdot \mathrm{IPM}_G(p_\Phi^{A=1}, p_\Phi^{A=0})\,. \tag{5}$$

The integral probability metric (IPM) (Müller, 1997; Sriperumbudur et al., 2012) measures the distance between two probability distributions p and q defined over M, *i.e.*, the latent space of R. Formally, $\mathrm{IPM}_G(p,q) \triangleq \sup_{g\in G} |\int_M g(m)\,(p(m)-q(m))\,dm|$, where $g: m \to \mathbb{R}$, represents a class of real-valued bounded measurable functions on M (Shalit et al., 2017). Therefore, model functions $\{h_a(\cdot), \Phi(\cdot)\}$ can be learned by minimizing the upper bound in (5) consisting of (i) only *factual* losses under both treatment assignments and (ii) an IPM regularizer enforcing latent distributional equivalence between the treatment groups. Note that if the data originates from a RCT it follows (by construction) that $\mathrm{IPM}_G(p_\Phi^{A=1}, p_\Phi^{A=0}) = 0$.

Accounting for censoring bias Below we formulate an approach for estimating functions $h_A(\cdot)$ and $\nu_A(\cdot)$ for synthesizing (sampling) non-censored $t_a \sim p_{h,\Phi}(T_A|X=x)$ and censored $c_a \sim p_{\nu,\Phi}(C_A|X=x)$ times, respectively. While some parametric assumptions for $p_{h,\Phi}(T_A|X=x)$ yield easy-to-evaluate closed forms for $S_{h,\Phi}(t_a|x)$ that can be used as likelihood for censored observations, they are restrictive, and have been shown to generate unrealistic high variance samples (Chapfuwa et al., 2018). So motivated, we seek a nonparametric likelihood-based approach that can model a flexible family of distributions, with an easy-to-sample approach for event times $t_a \sim p_{h,\Phi}(T_a|X=x)$. We conveniently model the event time generation process with a source of randomness, $p(\epsilon)$, *e.g.* Gaussian or uniform, which is obtained from a *planar flow* parameterized by $\{U_h, W_h, b_h\}$ (Rezende & Mohamed, 2015). Specifically, we transform the source of randomness, ϵ, using a single layer specification as follows

$$\tilde{\epsilon}_h = \epsilon + U_h \tanh(W_h\epsilon + b_h)\,, \quad \epsilon \sim \mathrm{Uniform}(0,1)\,, t_a = h_A(r, \tilde{\epsilon}_h)\,, \quad r = \Phi(x) \tag{6}$$

where $\{U_h, W_h\} \in \mathbb{R}^{d\times d}$, $\{b_h, \epsilon\} \in \mathbb{R}^d$, d is the dimensionality of the normalizing flow; each component of ϵ is drawn independently from $\mathrm{Uniform}(0,1)$, and $\tilde{\epsilon}_h$ may be viewed as a skip connection with stochasticity in ϵ. Further, $h_A(r, \tilde{\epsilon}_h)$ and $\Phi(x)$ are time-to-event generative and encoding functions, respectively, parameterized as neural networks. For simplicity, the dimensions of r and ϵ are set to d, however, they can be set independently if desired. In practice, we are interested in generating realistic event-time samples; therefore, we account for both censored and non-censored observations by adopting the objective from Chapfuwa et al. (2018), formulated as

$$\mathcal{L}_{\mathrm{F}}^{\mathrm{CSA}} \triangleq \mathbb{E}_{(y,\delta,x,a)\sim p(Y,\delta,X,A),\epsilon\sim p(\epsilon)}\left[\delta\cdot(|y-t_a|) + (1-\delta)\cdot(\max(0, y-t_a))\right]\,, \tag{7}$$

where the first term encourages sampled event times t_a to be close to y, the ground truth for observed events, *i.e.*, $\delta = 1$, while penalizing t_a for being smaller than the censoring time when $\delta = 0$. Further, the expectation is taken over samples (a minibatch) from empirical distribution $p(Y, \delta, X, A)$.

Informative censoring We model informative censoring similar to (7) but mirroring the censoring indicators to encourage accurate censoring time samples c_a for $\delta = 0$, while penalizing c_a for being smaller than y for $\delta = 1$ (observed events). Specifically, we set an independent planar flow like in (6) but parameterized by $\{U_\nu, W_\nu, b_\nu\}$ and censoring generative functions $\nu_A(r, \tilde{\epsilon}_\nu)$, parameterized as neural networks, where $c_a \sim p_{\nu,\Phi}(C_A|X=x)$ formulated as

$$\ell_c(\nu,\Phi) = \mathbb{E}_{(y,\delta,x,a)\sim p(y,\delta,X,A),\boldsymbol{\epsilon}\sim p(\epsilon)}\left[(1-\delta)\cdot(|y-c_a|) + \delta\cdot(\max(0, y-c_a))\right]\,. \tag{8}$$

Further, we introduce an additional time-order-consistency loss that enforces the correct order of the observed time relative to the censoring indicator, *i.e.*, $c_a < t_a$ if $\delta = 0$ and $t_a < c_a$ if $\delta = 1$, thus

$$\ell_{\mathrm{TC}}(h,\nu,\Phi) = \mathbb{E}_{(\delta,x,a)\sim p(\delta,X,A),\epsilon\sim p(\epsilon)}\left[\delta\cdot(\max(0, t_a - c_a)) + (1-\delta)\cdot(\max(0, c_a - t_a))\right] \tag{9}$$

Note that $\ell_{\mathrm{TC}}(h,\nu,\Phi)$ does not depend on the observed event times but only on the censoring indicators. Finally, we write the consolidated CSA loss for informative censoring (CSA-INFO) by aggregating (7), (8) and (9) as $\mathcal{L}_{\mathrm{F}}^{\mathrm{CSA-INFO}} \triangleq \mathcal{L}_{\mathrm{F}}^{\mathrm{CSA}} + \ell_{\mathrm{c}} + \ell_{\mathrm{TC}}$.

Learning Model functions $\{h_A(\cdot), \Phi(\cdot), \nu_A(\cdot)\}$ are learned by minimizing the bound (5), via stochastic gradient descent on minibatches from $\mathcal{D}$, with $\mathcal{L}_{\mathrm{F}}^{\mathrm{CSA}}$ for non-informative censoring and $\mathcal{L}_{\mathrm{F}}^{\mathrm{CSA-INFO}}$ for informative censoring. Further, for the IPM regularization loss in (5), we optimize the dual formulation of the *Wasserstein distance*, via the regularized *optimal transport* (Villani, 2008; Cuturi, 2013). Consequently, we only require $\alpha^{-1}\zeta_{h,\Phi}(x,a)$ to be 1-Lipschitz (Shalit et al., 2017) and α is selected by grid search on the validation set using *only* factual data (details below).

4 METRICS

We propose a comprehensive evaluation approach that accounts for both factual and causal metrics. Factual survival outcome predictions are evaluated according to standard survival metrics that measure diverse performance characteristics, such as concordance index (C-Index) (Harrell Jr et al., 1984), mean coefficient of variation (COV) and calibration slope (C-slope) (Chapfuwa et al., 2019). See the SM for more details on these metrics. For causal metrics, defined below, we introduce a nonparametric hazard ratio (HR) between treatment outcomes, and adopt the conventional precision in estimation of heterogeneous effect (PEHE) and average treatment effect (ATE) performance metrics (Hill, 2011). Note that PEHE and ATE require ground truth counterfactual event times, which is only possible in (semi-)synthetic data. For HR, we compare our findings with those independently reported in the literature from gold-standard RCT data.

Nonparametric Hazard Ratio In a medical setting, the population hazard ratio $\mathrm{HR}(t)$ between treatment groups is considered informative thus has been widely used in drug development and RCT (Yusuf et al., 2016; Mihaylova et al., 2012). For example, $\mathrm{HR}(t) < 1, > 1$, or ≈ 1 indicate *population* positive, negative and neutral treatment effects at time t, respectively. Moreover, $\mathrm{HR}(t)$ naturally accounts for both censored and non-censored outcomes. Standard approaches for computing $\mathrm{HR}(t)$ rely on the restrictive proportional hazard assumption from CoxPH (Cox, 1972), which is constituted as a semi-parametric linear model $\lambda(t|a) = \lambda_{\mathrm{b}}(t)\exp(a\beta)$. However, the constant covariate (time independent) effect is often violated in practice (see Figure 2b). For CoxPH, the *marginal* HR between treatment and control can be obtained from regression coefficient β learned via maximum likelihood without the need for specifying the baseline hazard $\lambda_{\mathrm{b}}(t)$: $\mathrm{HR}_{\mathrm{CoxPH}}(t) = \frac{\lambda(t|a=1)}{\lambda(t|a=0)} = \exp(\beta)$. So motivated, we propose a nonparametric, model-free approach for computing $\mathrm{HR}(t)$, in which we do not assume a parametric form for the event time distribution or the proportional hazard assumption from CoxPH. This approach only relies on samples from the conditional event time density functions, $f(t_1|x)$ and $f(t_0|x)$, via $t_a = h_A(\cdot)$ from (6).

Definition 1 *We define the nonparametric marginal Hazard Ratio and its approximation,* $\hat{\mathrm{HR}}(t)$*, as*

$$\mathrm{HR}(t) = \frac{\lambda_1(t)}{\lambda_0(t)} = \frac{S_0(t)}{S_1(t)} \cdot \frac{S_1'(t)}{S_0'(t)}, \quad \hat{\mathrm{HR}}(t) = \frac{\hat{S}_0^{\mathrm{PKM}}(t)}{\hat{S}_1^{\mathrm{PKM}}(t)} \cdot \frac{m_1(t)}{m_0(t)}, \tag{10}$$

where for $\mathrm{HR}(t)$ we leveraged $f(t|x) = \lambda(t|x)S(t|x)$ to obtain (10) and $S'(t) \triangleq dS(t)/dt$. The nonparametric assumption for $S(t)$ makes the computation of $S'(t)$ challenging. Provided that $S(t)$ is a monotonically decreasing function, for simplicity, we fit a linear function $S(t) = m \cdot t + c$, and set $S'(t) \approx m$. Note that the linear model is *only* used for estimating $S'(t)$ from the nonparametric estimation of $S(t)$. Bias from $S'(t)$ can be reduced by considering more complex function approximations for $S(t)$, *e.g.*, polynomial or spline. For the nonparametric estimation of $S(t)$ we leverage the *model-free* population point-estimate-based nonparametric Kaplan-Meier (Kaplan & Meier, 1958) estimator of the survival function $\hat{S}^{\mathrm{PKM}}(t)$ in Chapfuwa et al. (2019) to marginalize both *factual* and *counterfactual* predictions given covariates x. The approximated hazard ratio, $\hat{\mathrm{HR}}(t)$, is thus obtained by combining the approximations $\hat{S}_a^{\mathrm{PKM}}(t)$ and m_a. A similar formulation for the conditional, $\hat{\mathrm{HR}}(t|x)$, can also be derived. See the SM for full details on the evaluation or $\hat{\mathrm{HR}}(t)$ and $\hat{\mathrm{HR}}(t|x)$. Note that for some AFT- or CoxPH-based parametric formulations, $\mathrm{HR}(t|x)$, can be readily evaluated because $f(t_a|x)$ and $S(t_a|x)$ are available in closed form.

In the experiments, we will use $\mathrm{HR}(t)$ to compare different approaches against results reported in RCTs (see Tables 1 and 3). Further, we will use $\mathrm{HR}(t|x)$ to illustrate *stratified* treatment effects (see Figure 2). Note that though a neural-based survival recommender system (Katzman et al., 2018) has been previously used to estimate $\mathrm{HR}(t|x)$, their approach does not account for confounding or informative censoring thus it is susceptible to bias.

Precision in Estimation of Heterogeneous Effect (PEHE) A general *individualized* estimation error is formulated as $\epsilon_{\mathrm{PEHE}} = \sqrt{\mathbb{E}_X[(\mathrm{ITE}(x) - \hat{\mathrm{ITE}}(x))^2]}$, where $\mathrm{ITE}(x)$ is the ground truth, $\hat{\mathrm{ITE}}(x) = \mathbb{E}_T\left[\gamma\left(T_1\right) - \gamma\left(T_0\right)|X = x\right]$ and $\gamma(\cdot)$ is a deterministic transformation. In our experiments, $\gamma(\cdot)$ is the average over samples from $t_a \sim p_{h,\Phi}(T_A|X = x)$. Alternative estimands, *e.g.*, thresholding survival times $\gamma(T_A) = I\{T_A > \tau\}$, can also be considered as described above.

Average Treatment Effect (ATE) The *population* treatment effect estimation error is defined as $\epsilon_{\mathrm{ATE}} = |\mathrm{ATE} - \hat{\mathrm{ATE}}|$, where $\mathrm{ATE} = \mathbb{E}_X[\mathrm{ITE}(x)]$ (ground truth) and $\hat{\mathrm{ATE}} = \mathbb{E}_X[\hat{\mathrm{ITE}}(x)]$.

Table 1: Performance comparisons on ACTG-SYNTHETIC data, with 95% $\mathrm{HR}(t)$ confidence interval. The ground truth, test set, hazard ratio is $\mathrm{HR(t)} = 0.52_{(0.39,0.71)}$.

Method	Causal			Factual		
	ϵ_{PEHE}	ϵ_{ATE}	$\mathrm{HR}(t)$	C-Index (A=0, A=1)	Mean COV	C-Slope (A=0, A=1)
CoxPH-Uniform	NA	NA	$0.97_{(0.86,1.09)}$	NA	NA	NA
CoxPH-IPW	NA	NA	$0.48_{(0.03,7.21)}$	NA	NA	NA
CoxPH-OW	NA	NA	$0.60_{(0.53,0.68)}$	NA	NA	NA
Surv-BART	352.07	77.89	$0.0(0.0, 0.0)$	(0.706, 0.686)	0.001	(0.398, ∞)
AFT-Weibull	367.92	133.93	$0.47_{(0.47,0.47)}$	(0.21, 0.267)	6.209	(0.707, 0.729)
AFT-log-Normal	377.76	157.64	$0.47_{(0.47,0.47)}$	(0.675, 0.556)	6.971	(0.707, 0.729)
SR	369.47	88.55	$0.38_{(0.33,0.65)}$	(0.791, 0.744)	0	(0.985, 1.027)
CSA (proposed)	358.72	**0.8**	$0.45_{(0.39,0.65)}$	(0.787, 0.767)	0.131	(0.985, 1.026)
CSA-INFO (proposed)	**344.3**	31.19	$\mathbf{0.53}_{(0.41,0.67)}$	(0.78, 0.764)	0.13	(0.999, 1.029)

5 EXPERIMENTS

We describe the baselines and datasets that will be used to evaluate the proposed counterfactual survival analysis methods (CSA and CSA-INFO). Pytorch code including the new semi-synthetic dataset (see below) will be made publicly available. Throughout the experiments, we use the standard $\mathrm{HR}(t)$ for CoxPH based methods and (10) for all others. The bound in (5) is sensitive to α, thus we propose approximating *proxy* counterfactual outcomes $\{Y_{\text{CF}}, \delta_{\text{CF}}\}$ for the validation set, according to the covariate Euclidean nearest-neighbour (NN) from the training set. We select the α that minimizes the validation loss $\mathcal{L} = \mathcal{L}_{\text{F}} + \mathcal{L}_{\text{CF}}$ from the set $(0, 0.1, 1, 10, 100)$.

Baselines We consider the following competitive baseline approaches: (i) *propensity* weighted CoxPH (Schemper et al., 2009; Buchanan et al., 2014; Rosenbaum & Rubin, 1983); (ii) IPM (5) regularized AFT (log-Normal and Weibull) models; (iii) an IPM (5) regularized *deterministic* semi-supervised regression (SR) model with accuracy objective from (Chapfuwa et al., 2018), as a contrast for the proposed stochastic predictors (CSA and CSA-INFO); and (iv) survival Bayesian additive regression trees (Surv-BART) (Sparapani et al., 2016). For CoxPH, we consider three normalized weighting schemes: (i) inverse probability weighting (IPW) (Horvitz & Thompson, 1952; Cao et al., 2009), where $\mathrm{IPW}_i = \frac{a_i}{\hat{e}_i} + \frac{1-a_i}{1-\hat{e}_i}$; $ii)$ overlapping weights (OW) (Crump et al., 2006; Li et al., 2018), where $\mathrm{OW}_i = a_i \cdot (1 - \hat{e}_i) + (1 - a_i) \cdot \hat{e}_i$; and $iii)$ the standard RCT uniform assumption. A simple linear logistic model $\hat{e}_i = \sigma(x_i; w)$, is used as an approximation, $\hat{e}_i$, to the unknown propensity score $P(A = 1|X = x)$. See the SM for a details of the baselines.

Datasets We consider the following datasets: (i) FRAMINGHAM, is an EHR-based longitudinal cardiovascular cohort study that we use to evaluate the effect of statins on future coronary heart disease outcomes (Benjamin et al., 1994); (ii) ACTG, is a longitudinal RCT study comparing monotherapy with Zidovudine or Didanosine with combination therapy in HIV patients (Hammer et al., 1996); and (iii) ACTG-SYNTHETIC, is a semi-synthetic dataset based on ACTG covariates. We simulate potential outcomes according to a Gompertz-Cox distribution (Bender et al., 2005) with selection bias from a simple logistic model for $P(A = 1|X = x)$ and AFT-based censoring mechanism. The generative process is detailed in the SM. Table 2 summarizes the datasets according to (i) covariates of size p; (ii) proportion of non-censored events, treated units, and missing entries in the $N \times p$ covariate matrix; and (iii) time range $t_{\max}$ for both censored and non-censored events. Missing entries are imputed with median or mode if continuous or categorical, respectively.

Table 2: Summary statistics of the datasets.

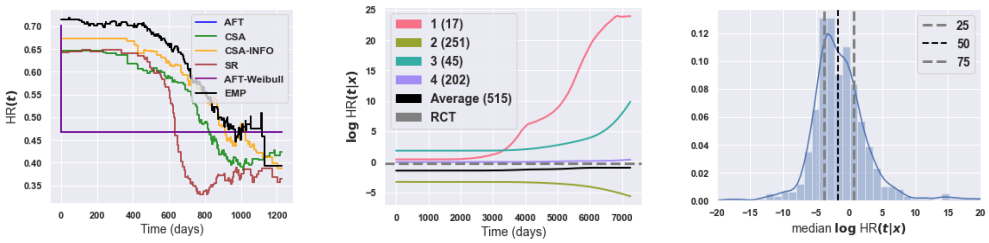
	FRAMINGHAM	ACTG	ACTG-SYNTHETIC
Events (%)	26.0	26.9	48.9
Treatment (%)	10.4	49.5	55.9
N	3,435	1,054	2,139
p	32	23	23
Missing (%)	0.23	1.41	1.38
$t_{\max}$ (days)	7,279	1,231	1,313

Quantitative Results Experimental results for two data-sets in Tables 1 and 3, illustrate that AFT-based methods are high variance, inferior in calibration and C-Index than accuracy-based methods (SR, CSA, CSA-INFO). Surv-BART is the least calibrated but low variance method. CSA-INFO and CSA outperform all methods across all factual metrics, whereas CSA-INFO is better calibrated, low variance but slightly lower C-Index than CSA. Note that we fit CoxPH using the entire dataset; since it does not support counterfactual inference, we do not present factual metrics. By properly adjusting for both informative censoring and selection bias, CSA-INFO significantly outperforms all methods in treatment effect estimation according to $\mathrm{HR}(t)$ and ϵ_{PEHE}, across non-RCT datasets, while remaining comparable to AFT-Weibull on the RCT dataset (see the SM). Further, RCT-based results on ACTG data in the SM illustrate comparable $\mathrm{HR}(t)$ across all models except for AFT-log-Normal and Surv-BART, which overestimate, and SR, which underestimates risk. For non-RCT datasets (ACTG-SYNTHENTIC and FRAMINGHAM), CoxPH-OW has a clear advantage over all CoxPH based

Table 3: Performance comparisons on FRAMINGHAM data, with 95% $\text{HR}(t)$ confidence interval. Test set NN assignment of y_{CF} and δ_{CF} yields biased $\text{HR(t)} = 1.23_{(1.17,1.25)}$, while previous large scale longitudinal RCT studies estimated $\text{HR(t)} = 0.75_{(0.64,0.88)}$ (Yusuf et al., 2016).

Method	Causal	Factual		
	$\text{HR}(t)$	C-Index (A=0, A=1)	Mean COV	C-Slope (A=0, A=1)
CoxPH-Uniform	$1.69_{(1.38,2.07)}$	NA	NA	NA
CoxPH-IPW	$1.09_{(0.76,1.57)}$	NA	NA	NA
CoxPH-OW	$0.88_{(0.73,1.08)}$	NA	NA	NA
Surv-BART	$14.99_{(14.9,14.9e8)}$	(0.629, 0.630)	0.003	(0.232, 0.084)
AFT-Weibull	$1.09_{(1.09,1.09)}$	(0.734, 0.395)	8.609	(0.857, 0.89)
AFT-log-Normal	$1.55_{(1.46,1.55)}$	(0.68, 0.56)	10.415	(0.979, 0.732)
SR	$0.58_{(0.53,0.71)}$	(0.601, 0.57)	0	(0.491, 0.63)
CSA (proposed)	$1.04_{(1.00,1.09)}$	(0.763, 0.728)	0.161	(0.891, 0.81)
CSA-INFO (proposed)	$\mathbf{0.81}_{(0.77,0.83)}$	(0.752, 0.651)	0.156	(0.907, 0.881)

methods, mostly credited to the well-behaved bounded propensity weights $\in [0, 1]$. Interestingly, the FRAMINGHAM observational data exhibits a common paradox, where without proper adjustment of selection and censoring bias, naive approaches would result in a counter-intuitive treatment effect from statins. However, there is severe *confounding* from covariates such as age, BMI, diabetes, CAD, PAD, MI, stroke, *etc*., that influence both treatment likelihood and survival time. Table 3, demonstrates that CSA-INFO is clearly the best performing approach. Specifically, its $\text{HR}(t)$, reverses the biased observational treatment effect, to demonstrate positive treatment from statins, which is consistent with prior large RCT longitudinal findings (Yusuf et al., 2016).



(a) ACTG-SYNTHENTIC $\text{HR}(t)$ (b) FRAMINGHAM $\log \text{HR}(t|x)$ (c) FRAMINGHAM $\log \text{HR}(t|x)$ pdf

Figure 2: (a) Inferred population $\text{HR}(t)$ compared against ground truth (EMP) on ACTG-SYNTHETIC data. CSA-INFO-based (b) cluster-specific average $\log \text{HR}(t|x)$ curves and (c) estimated density of median $\log \text{HR}(t|x)$ values on the test set of the FRAMINGHAM dataset. Clusters assignment were obtained via hierarchical clustering of individualized $\log \text{HR}(t|x)$ traces.

Qualitative Results Figure 2a demonstrates that CSA-INFO matches the ground truth population hazard, $\text{HR}(t)$, better than alternative methods on ACTG-SYNTHETIC data. See the SM for ACTG and FRAMINGHAM. Figure 2b shows sub-population log hazard ratios for four patient clusters obtained via hierarchical clustering on the individual log hazard ratios, $\log \text{HR}(t|x)$, of the test set of FRAMINGHAM data. Interestingly, these clusters stratify treatment effects into: positive (2), negative (1 and 3), and neutral (4) sub-populations. Moreover, the estimated density of median $\log \text{HR}(t|x)$ values in Figure 2c illustrates that nearly 70% of the testing set individuals have $\log \text{HR}(t|x) < 0$, thus may benefit from taking statins. Further, we isolated the extreme top and bottom quantiles, $\text{HR}(t|x) < 0.024$ and $\text{HR}(t|x) > 1.916$, respectively, of the median $\log \text{HR}(t|x)$ values for the test set of FRAMINGHAM, as shown in Figure 2c. After comparing their covariates, we found that individuals with the following characteristics may benefit from taking statins: young, male, diabetic, without prior history (CAD, PAD, stroke or MI), high BMI, cholesterol, triglycerides, fasting glucose, and low high-density lipoprotein. There seem to be consensus that diabetics and high-cholesterol patients benefit from statins (Cheung et al., 2004; Wilt et al., 2004). See SM for additional results.

6 CONCLUSIONS

We have proposed a unified counterfactual inference framework for survival analysis. Our approach adjusts for bias from two unknown sources, namely, *confounding* due to covariate dependent selection bias and *censoring* (informative or non-informative). Relative to competitive alternatives, we demonstrate superior performance for both survival-outcome prediction and treatment-effect estimation, across three diverse datasets, including a semi-synthetic dataset which we introduce. Moreover, we formulate a model-free nonparametric hazard ratio metric for comparing treatment effects or leveraging prior randomized real-world experiments in longitudinal studies.

REFERENCES

Peter C Austin. Propensity-score matching in the cardiovascular surgery literature from 2004 to 2006: a systematic review and suggestions for improvement. *The Journal of Thoracic and Cardiovascular Surgery*, 2007.

Peter C Austin. The use of propensity score methods with survival or time-to-event outcomes: reporting measures of effect similar to those used in randomized experiments. *Statistics in Medicine*, 2014.

Elias Bareinboim and Judea Pearl. Controlling selection bias in causal inference. In *AISTATS*, 2012.

Ralf Bender, Thomas Augustin, and Maria Blettner. Generating survival times to simulate cox proportional hazards models. *Statistics in medicine*, 2005.

Emelia J Benjamin, Daniel Levy, Sonya M Vaziri, Ralph B D'Agostino, Albert J Belanger, and Philip A Wolf. Independent risk factors for atrial fibrillation in a population-based cohort: the framingham heart study. *Jama*, 1994.

Ashley L Buchanan, Michael G Hudgens, Stephen R Cole, Bryan Lau, Adaora A Adimora, and Women's Interagency HIV Study. Worth the weight: using inverse probability weighted cox models in aids research. *AIDS research and human retroviruses*, 2014.

Weihua Cao, Anastasios A Tsiatis, and Marie Davidian. Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*, 2009.

Paidamoyo Chapfuwa, Chenyang Tao, Chunyuan Li, Courtney Page, Benjamin Goldstein, Lawrence Carin, and Ricardo Henao. Adversarial time-to-event modeling. In *ICML*, 2018.

Paidamoyo Chapfuwa, Chenyang Tao, Lawrence Carin, and Ricardo Henao. Survival function matching for calibrated time-to-event predictions. *arXiv*, 2019.

Bernard MY Cheung, Ian J Lauder, Chu-Pak Lau, and Cyrus R Kumana. Meta-analysis of large randomized controlled trials to evaluate the impact of statins on cardiovascular outcomes. *British journal of clinical pharmacology*, 2004.

Hugh A Chipman, Edward I George, Robert E McCulloch, et al. Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 2010.

Stephen R Cole and Miguel A Hernán. Adjusted survival curves with inverse probability weights. *Computer methods and programs in biomedicine*, 2004.

David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 1972.

Richard K Crump, V Joseph Hotz, Guido W Imbens, and Oscar A Mitnik. Moving the goalposts: Addressing limited overlap in the estimation of average treatment effects by changing the estimand. Technical report, National Bureau of Economic Research, 2006.

Yifan Cui, Michael R Kosorok, Stefan Wager, and Ruoqing Zhu. Estimating heterogeneous treatment effects with right-censored data via causal survival forests. *arXiv*, 2020.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *NeurIPS*, 2013.

Iván Díaz. Statistical inference for data-adaptive doubly robust estimators with survival outcomes. *Statistics in Medicine*, 2019.

Jennifer Frankovich, Christopher A Longhurst, and Scott M Sutherland. Evidence-based medicine in the emr era. *N Engl J Med*, 2011.

Saurabh Gombar, Alison Callahan, Robert Califf, Robert Harrington, and Nigam H Shah. It is time to learn from patients like mine. *NPJ digital medicine*, 2019.

Scott M Hammer, David A Katzenstein, Michael D Hughes, Holly Gundacker, Robert T Schooley, Richard H Haubrich, W Keith Henry, Michael M Lederman, John P Phair, Manette Niu, et al. A trial comparing nucleoside monotherapy with combination therapy in hiv-infected adults with cd4 cell counts from 200 to 500 per cubic millimeter. *New England Journal of Medicine*, 1996.

Frank E Harrell Jr, Kerry L Lee, Robert M Califf, David B Pryor, and Robert A Rosati. Regression modelling strategies for improved prognostic prediction. *Statistics in medicine*, 1984.

Kristiina Häyrinen, Kaija Saranto, and Pirkko Nykänen. Definition, structure, content, use and impacts of electronic health records: a review of the research literature. *International Journal of Medical Informatics*, 2008.

Nicholas C Henderson, Thomas A Louis, Gary L Rosner, and Ravi Varadhan. Individualized treatment effects with censored data via fully nonparametric bayesian accelerated failure time models. *Biostatistics*, 2020.

Miguel A Hernán and James M Robins. Causal inference: what if. *Boca Raton: Chapman & Hill/CRC*, 2020.

Miguel A Hernán, Stephen R Cole, Joseph Margolick, Mardge Cohen, and James M Robins. Structural accelerated failure time models for survival analysis in studies with time-varying treatments. *Pharmacoepidemiology and Drug Safety*, 2005.

Miguel Ángel Hernán, Babette Brumback, and James M Robins. Marginal structural models to estimate the causal effect of zidovudine on the survival of hiv-positive men. *Epidemiology*, 2000.

Jennifer L Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 2011.

Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 1952.

Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, Michael S Lauer, et al. Random survival forests. *The annals of applied statistics*, 2008.

Ashish K Jha, Catherine M DesRoches, Eric G Campbell, Karen Donelan, Sowmya R Rao, Timothy G Ferris, Alexandra Shields, Sara Rosenbaum, and David Blumenthal. Use of electronic health records in us hospitals. *New England Journal of Medicine*, 2009.

Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 1958.

Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*, 2018.

David G Kleinbaum and Mitchel Klein. *Survival analysis*. Springer, 2010.

Fan Li, Kari Lock Morgan, and Alan M Zaslavsky. Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 2018.

Christopher A Longhurst, Robert A Harrington, and Nigam H Shah. A 'green button'for using aggregate patient data at the point of care. *Health affairs*, 2014.

B Mihaylova, J Emberson, L Blackwell, A Keech, J Simes, EH Barnes, M Voysey, 3A Gray, R Collins, and C Baigent. The effects of lowering ldl cholesterol with statin therapy in people at low risk of vascular disease: meta-analysis of individual data from 27 randomised trials., 2012.

Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 1997.

Judea Pearl and Elias Bareinboim. External validity: From do-calculus to transportability across populations. *Statistical Science*, 2014.

Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *ICML*, 2015.

James Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, 1986.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 1983.

Donald B Rubin. Causal inference using potential outcomes. *Journal of the American Statistical Association*, 2005.

Michael Schemper, Samo Wakounig, and Georg Heinze. The estimation of average hazard ratios by weighted cox regression. *Statistics in medicine*, 2009.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *ICML*, 2017.

Jincheng Shen, Lu Wang, Stephanie Daignault, Daniel E Spratt, Todd M Morgan, and Jeremy MG Taylor. Estimating the optimal personalized treatment strategy based on selected variables to prolong survival via random survival forest with weighted bootstrap. *Journal of biopharmaceutical statistics*, 2018.

Rodney A Sparapani, Brent R Logan, Robert E McCulloch, and Purushottam W Laud. Nonparametric survival analysis using bayesian additive regression trees (bart). *Statistics in medicine*, 2016.

Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, Gert RG Lanckriet, et al. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 2012.

Ludovic Trinquart, Justine Jacot, Sarah C Conner, and Raphaël Porcher. Comparison of treatment effects measured by the hazard ratio and by the ratio of restricted mean survival times in oncology randomized controlled trials. *Journal of Clinical Oncology*, 2016.

Anastasios Tsiatis. *Semiparametric theory and missing data*. Springer Science & Business Media, 2007.

Mark J van der Laan and James M Robins. Unified approach for causal inference and censored data. In *Unified Methods for Censored Longitudinal Data and Causality*. Springer, 2003.

Mark J Van der Laan and Sherri Rose. *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media, 2011.

Cédric Villani. *Optimal transport: old and new*. Springer Science & Business Media, 2008.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 2018.

Lee-Jen Wei. The accelerated failure time model: a useful alternative to the cox regression model in survival analysis. *Statistics in medicine*, 1992.

Timothy J Wilt, Hanna E Bloomfield, Roderick MacDonald, David Nelson, Indulis Rutks, Michael Ho, Gregory Larsen, Anthony McCall, Sandra Pineros, and Anne Sales. Effectiveness of statin therapy in adults with coronary heart disease. *Archives of internal medicine*, 2004.

Salim Yusuf, Jackie Bosch, Gilles Dagenais, Jun Zhu, Denis Xavier, Lisheng Liu, Prem Pais, Patricio López-Jaramillo, Lawrence A Leiter, Antonio Dans, et al. Cholesterol lowering in intermediate-risk persons without cardiovascular disease. *New England Journal of Medicine*, 2016.

Yao Zhang, Alexis Bellot, and Mihaela van der Schaar. Learning overlapping representations for the estimation of individualized treatment effects. In *AISTATS*, 2020.

Lihui Zhao, Lu Tian, Hajime Uno, Scott D Solomon, Marc A Pfeffer, Jerald S Schindler, and Lee Jen Wei. Utilizing the integrated difference of two survival functions to quantify the treatment contrast for designing, monitoring, and analyzing a comparative clinical study. *Clinical trials*, 2012.