Bioinformatics, 38, 2022, i60–i67
https://doi.org/10.1093/bioinformatics/btac221
ISCB/ISMB 2022

BITES: balanced individual treatment effect for survival data

S. Schrod[1,*], A. Schäfer[2], S. Solbrig[2], R. Lohmayer[3], W. Gronwald[4], P. J. Oefner[4], T. Beißbarth[1], R. Spang[5], H. U. Zacharias[6,7], and M. Altenbuchinger[1,*]

[1]Department of Medical Bioinformatics, University Medical Center Göttingen, Göttingen 37077, Germany, [2]Department of Physics, Institute of Theoretical Physics, University of Regensburg, Regensburg 93051, Germany, [3]Leibniz Institute for Immunotherapy, Regensburg 93053, Germany, [4]Institute of Functional Genomics, University of Regensburg, Regensburg 93053, Germany, [5]Department of Statistical Bioinformatics, Institute of Functional Genomics, University of Regensburg, Regensburg 93053, Germany, [6]Department of Internal Medicine I, University Medical Center Schleswig-Holstein, Campus Kiel, Kiel 24105, Germany and [7]Institute of Clinical Molecular Biology, Kiel University and University Medical Center Schleswig-Holstein, Campus Kiel, Kiel 24105, Germany

*To whom correspondence should be addressed.

Abstract

Motivation: Estimating the effects of interventions on patient outcome is one of the key aspects of personalized medicine. Their inference is often challenged by the fact that the training data comprises only the outcome for the administered treatment, and not for alternative treatments (the so-called counterfactual outcomes). Several methods were suggested for this scenario based on observational data, i.e. data where the intervention was not applied randomly, for both continuous and binary outcome variables. However, patient outcome is often recorded in terms of time-to-event data, comprising right-censored event times if an event does not occur within the observation period. Albeit their enormous importance, time-to-event data are rarely used for treatment optimization. We suggest an approach named BITES (Balanced Individual Treatment Effect for Survival data), which combines a treatment-specific semi-parametric Cox loss with a treatment-balanced deep neural network; i.e. we regularize differences between treated and non-treated patients using Integral Probability Metrics (IPM).

Results: We show in simulation studies that this approach outperforms the state of the art. Furthermore, we demonstrate in an application to a cohort of breast cancer patients that hormone treatment can be optimized based on six routine parameters. We successfully validated this finding in an independent cohort.

Availability and implementation: We provide BITES as an easy-to-use python implementation including scheduled hyper-parameter optimization (https://github.com/sschrod/BITES). The data underlying this article are available in the CRAN repository at https://rdrr.io/cran/survival/man/gbsg.html and https://rdrr.io/cran/survival/man/rotterdam.html.

Contact: stefan.schrod@bioinf.med.uni-goettingen.de or michael.altenbuchinger@bioinf.med.uni-goettingen.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Inferring the effect of interventions on outcomes is relevant in diverse domains, comprising precision medicine and epidemiology (Frieden, 2017) or marketing (Bottou *et al.*, 2013; Kohavi *et al.*, 2009). A fundamental issue of causal reasoning is that potential outcomes are observed only for the applied intervention but not for its alternatives (the counterfactuals). This is particularly true in medicine, where patient's outcome is only known for the applied (the factual) treatment and not for its alternatives, i.e. the counterfactual outcomes remain hidden. Therapeutic interventions, such as drug treatments or surgeries, are typically made by physicians on the basis of expert consensus guidelines. This process has to take into account both the expected treatment benefit, but also the potential side effects. Estimates of the former can be difficult. For instance, the success of drug treatments in cancer strongly depends on multiple characteristics of the tumor and the patient, and, consequently, the Average Treatment Effect (ATE) estimated from controlled trials does not necessarily hold on the level of individual patients. Thus, an estimate of the Individual Treatment Effect (ITE) is necessary, which has to be inferred from data (Holland, 1986). Solving the latter 'missing data problem' was attempted repeatedly in the literature using machine learning methods in combination with counterfactual reasoning. There are two naive approaches to this issue: the treatment can be included as a covariate or it can be used to stratify the model development, i.e. individual treatment-specific models are learned (also called T-learner). Potential outcomes can then be estimated by changing the respective treatment covariate or model. These naive approaches are occasionally discussed in performance comparisons, e.g. in Chapfuwa *et al.*, (2020) and Curth *et al.* (2021). An alternative approach is to match similar patients between treated and non-treated populations using, e.g. propensity scores (Rosenbaum and Rubin, 1983). This directly provides estimates of counterfactual outcomes. However, a central issue in this context is

to define appropriate similarity measures, which should ideally also be valid in a high-dimensional variable space (King and Nielsen, 2019). Further alternatives are Causal Forests (Athey *et al.*, 2016; Athey and Wager, 2019; Wager and Athey, 2017) or deep architectures such as the Treatment-Agnostic Representation Network (TARNet) (Johansson *et al.*, 2016; Shalit *et al.*, 2016). Both methods do not account for treatment selection biases and thus will be biased toward treatment-specific distributions. This issue was recently approached by several groups which balanced the treated and non-treated distributions using model regularization via representations of Integral Probability Metrics (IPM) (Müller, 1991). Suggested methods are, e.g. balanced propensity score matching (Diamond and Sekhon, 2013; Li and Fu, 2017), deep implementations such as the Counterfactual regression Network (CFRNet) (Johansson *et al.*, 2016; Shalit *et al.*, 2016) or the auto-encoder based Deep-Treat (Atan *et al.*, 2018). Recently, balancing was incorporated in a Generative Adversarial Net for inference of Individualized Treatment Effects (GANITE) (Yoon *et al.*, 2018). Note, learning balanced representations involves a trade-off between predictive power and bias since biased information can be also highly predictive.

All aforementioned approaches deal with continuous or binary response variables. In medicine, however, patient outcome is often recorded as time-to-event data, i.e. the time until an event occurs. The patient is (right-)censored at the last known follow-up if the event was not observed within the observation period. A plethora of statistical approaches deal with the analysis of time-to-event data (Martinussen and Scheike, 2006), of which one of the most popular methods is Cox's Proportional Hazards (PH) model (Cox, 1972). The Cox PH model is a semi-parametric approach for time-to-event data, which models the influence of variables on the baseline hazard. Here, the PH assumption implies an equal baseline hazard for all observations. In fact, the influence of variables can be estimated without any consideration of the baseline hazard function (Breslow, 1972; Cox, 1972). The Cox PH model is also highly relevant in the context of machine learning. It was adapted to the high-dimensional setting using l_1 and l_2 regularization terms (Tibshirani, 1997), with applications ranging from the prediction of adverse events in patients with chronic kidney disease (Zacharias *et al.*, 2021) to the risk prediction in cancer entities (Jachimowicz *et al.*, 2021; Staiger *et al.*, 2020). The Cox PH model can be also adapted to deep learning architectures, as proposed by (Katzman *et al.*, 2018). Alternative machine-learning approaches to model time-to-event data include discrete-time Cox models built on multi-outcome feedforward architectures (Gensheimer and Narasimhan, 2019; Kvamme and Borgan, 2019; Lee *et al.*, 2018) and random survival forests (RSF) (Athey and Wager, 2019; Ishwaran *et al.*, 2008).

The prediction of ITEs from time-to-event data has received little attention in the machine learning community, which is surprising considering the enormous practical relevance of the topic. Seminal works are (Chapfuwa *et al.*, 2020) and (Curth *et al.*, 2021). Most recently, Curth *et al.* (2021) suggested to learn discrete-time treatment-specific conditional hazard functions, which were estimated using a deep learning approach. Treatment and control distributions were balanced analogously to Shalit *et al.* (2016) using the p-Wasserstein distance (Kantorovitch, 1958; Ramdas *et al.*, 2017). This approach, named SurvITE, was shown to outperform the current state of the art in simulation studies.

We propose to combine the loss of the Cox PH model with an IPM regularized deep neural network architecture to balance generating distributions of treated and non-treated patients. We named this approach 'Balanced Individual Treatment Effect for Survival data' (BITES). We show that this approach—albeit its apparent simplicity—outcompetes SurvITE as well as alternative state-of-the-art methods. First, we demonstrate the superior performance of BITES in simulation studies where we focus on biased treatment assignments and small sample sizes. Second, we used training data from the Rotterdam Tumour Bank (Foekens *et al.*, 2000) to show that BITES can optimize hormone treatment in patients with breast cancer. We validated the latter model using data from a controlled randomized trial of the German Breast Cancer Study Group (GBSG) (Schumacher *et al.*, 1994) and analyzed feature importance using SHAP (SHapley Additive exPlanations) values (Lundberg and Lee, 2017). We further provide an easy-to-use python implementation of BITES including scheduled hyper-parameter optimization (https://github.com/sschrod/BITES).

2 Materials and methods

Patient outcome can be recorded as (right-)censored time-to-event data. First, we will introduce models for such data, i.e. the Cox PH model and recent non-linear adaptations. Second, we will discuss the potential outcome model and how it can be used to model survival. Third, we introduce regularization techniques to account for unbalanced distributions and, finally, we will combine these methods in a deep neural network approach termed BITES to learn treatment recommender systems based on patient survival.

2.1 Survival data

Let $\mathcal{X}$ be the space of covariates and $\mathcal{T}$ the space of available treatments. Furthermore, let $y \in \mathcal{Y}$ be the observed survival times and $E \in \mathcal{E} = \{0, 1\}$ the corresponding event indicator. Denote sample data of patient i by the triplet $(\boldsymbol{x_i}, y_i, E_i) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{E}$. If the patient experiences the event within the observation period, $y_i^{E=1}$ is the time until the event of interest occurs, otherwise $y_i^{E=0}$ is the censoring time. Let the survival times y be distributed according to $f(y)$ with the corresponding cumulated event distribution $F(y) = \int_0^y f(y')\mathrm{d}y'$. The survival probability at time y is then given by $S(y) = 1 - F(y)$. The hazard function is

$$\lambda(y; \boldsymbol{x}) = \underbrace{\exp(\boldsymbol{\beta}^T \boldsymbol{x})}_{\text{hazard rate}} \lambda_0(y) \tag{1}$$

and corresponds to the risk of dying at time y (Cox, 1972), i.e. a greater hazard corresponds to greater risk of failure. Here, the model parameters are given by $\boldsymbol{\beta}$ and the baseline hazard function is $\lambda_0(y) = \lambda(y; \boldsymbol{x} = 0)$. Note that $\lambda_0(y) = \frac{f(y)}{1-F(y)} = -\frac{\mathrm{d}}{\mathrm{d}y}\log(S(y))$. According to Cox's PH assumption, all patients share the same baseline hazard function and, importantly, the baseline hazard cancels in maximum likelihood estimates of $\boldsymbol{\beta}$. Thus, time dependence can be eliminated from the individual hazard prediction and rather than learning the exact time to event, Cox regression learns an ordering of hazard rates. At every event time $y_i^{E=1}$, the set of patients at risk is given by $\mathcal{R}_i = \mathcal{Y}(y \geq y_i^{E=1})$. The partial log-likelihood of the Cox model (Breslow, 1972; Cox, 1972) is given by:

$$\mathcal{L}(\boldsymbol{\beta}) = \sum_{i:E_i=1} \left[\log \left(\sum_{j:y_j \in R_i} e^{\boldsymbol{\beta}^T \boldsymbol{x}_j} \right) - \boldsymbol{\beta}^T \boldsymbol{x}_i \right]. \tag{2}$$

Faraggi and Simon (1995) suggested to replace the ordinary linear predictor function, $\boldsymbol{\beta}^T \boldsymbol{x}$, by a feedforward neural network with a single outcome node $h_\theta(\boldsymbol{x})$ and network parameters θ. Following this idea, Katzman *et al.* (2018) introduced DeepSurv, which showed improved performance compared to the linear case, particularly if non-linear covariate dependencies are present.

2.2 The counterfactual problem

The outcome space for multiple treatment options k is given by $\mathcal{Y} = \mathcal{Y}_0 \times \cdots \times \mathcal{Y}_{(k-1)}$. For simplicity, we will restrict the discussion to the binary case, $k = 2$, with a treated group, $T = 1$, and a control group, $T = 0$.

We consider the problem where only a single *factual* outcome is observed per patient, i.e. the outcomes for all other interventions, also known as the *counterfactuals*, are missing. Hence, the *individual treatment effect* (ITE), defined as

$$\tau(\boldsymbol{x_i}) = Y^{T=1}(\boldsymbol{x_i}) - Y^{T=0}(\boldsymbol{x_i}), \tag{3}$$

can only be inferred based on potential outcome estimates (Rubin, 1974). We will build a recommendation model that assigns treatments to patients with predictions $\tau(\boldsymbol{x_i}) > 0$.

Following recent work (Alaa and van der Schaar, 2017; Athey and Wager, 2019; Johansson *et al.*, 2016, 2020; Shalit *et al.*, 2016; Wager and Athey, 2017; Yao *et al.*, 2018; Yoon *et al.*, 2018), we make the standard *strong ignorability* assumption, which has been shown to be a sufficient condition to make the ITE identifiable (Pearl, 2017; Shalit *et al.*, 2016), i.e. it guarantees proper causal dependencies on the interventions. The *strong ignorability* assumption contains the *unconfoundedness* and *overlap* assumptions:

THEOREM 1 (Unconfoundedness). Covariates X do not simultaneously influence the treatment T and potential outcomes $(Y^{T=0}, Y^{T=1})$, i.e.

$$(Y^{T=0}, Y^{T=1}) \perp\!\!\!\perp T|X. \tag{4}$$

This assumption ensures that the causal effect is not influenced by non-observable causal substructures such as confounding (Pearl, 2009). Correcting for confounding bias requires structural causal models, which are ambiguous in general and need to be justified based on domain knowledge (Pearl, 2008).

THEOREM 2 (Overlap). There is a non-zero probability for each patient i to receive each of the treatments $T \in \mathcal{T}$:

$$0 < p(T_i|\boldsymbol{x}_i) < 1. \tag{5}$$

2.3 Balancing distributions

Strong ignorability only removes confounding artifacts. Imbalances of the generating distributions due to biased treatment administration might still be present. Balancing the generating distributions of treated and control group has been shown to be effective both on the covariate space (Imai and Ratkovic, 2014) and on latent representations (D'Amour *et al.*, 2017; Huang *et al.*, 2016; Johansson *et al.*, 2016, 2020; Li and Fu, 2017; Lu *et al.*, 2020; Shalit *et al.*, 2016; Yao *et al.*, 2018). This is either achieved by multi-task models or IPMs. The latter quantify the difference of probability measures $\mathbb{P}$ and $\mathbb{Q}$ defined on a measurable space S by finding a function $f \in \mathcal{F}$ that maximizes (Müller, 1991)

$$d_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) := \sup_{f \in \mathcal{F}} \left| \int f \, \mathrm{d}\mathbb{P} - \int f \, \mathrm{d}\mathbb{Q} \right|. \tag{6}$$

Most commonly used are the Maximum Mean Discrepancy (MMD), restricting the function space to reproducing kernel-Hilbert spaces (Gretton *et al.*, 2012), or the p-Wasserstein distance (Ramdas *et al.*, 2017). Both have appealing properties and can be empirically estimated (Sriperumbudur *et al.*, 2012). MMD has low sample complexity with a fast rate of convergence, which comes with low computational costs. A potential issue is that gradients vanish for overlapping means (Feydy *et al.*, 2018). The p-Wasserstein distance, on the other hand, offers more stable gradients even for overlapping means, which comes with higher computational costs, i.e. by solving a linear program. The computational burden can be reduced by entropically smoothing the latter and by using the Sinkhorn divergence,

$$S_\epsilon^p(\mathbb{P}, \mathbb{Q}) := \mathrm{W}_\epsilon^p(\mathbb{P}, \mathbb{Q}) - \frac{1}{2}\mathrm{W}_\epsilon^p(\mathbb{P}, \mathbb{P}) - \frac{1}{2}\mathrm{W}_\epsilon^p(\mathbb{Q}, \mathbb{Q}), \tag{7}$$

where $\mathrm{W}_\epsilon^p(\mathbb{P}, \mathbb{Q})$ is the smoothed Optimal Transport (OT) loss defined in the following (Feydy *et al.*, 2018; Ramdas *et al.*, 2017).

DEFINITION 1 (Smoothed Optimal Transport loss). For $p \in [1, \infty)$ and Borel probability measures $\mathbb{P}$, $\mathbb{Q}$ on $\mathbb{R}^d$ the entropically smoothed OT loss is

$$\begin{aligned} &\mathrm{W}_\epsilon^p(\mathbb{P}, \mathbb{Q}) := \min_{\pi \in \Gamma(\mathbb{P}, \mathbb{Q})} \int_{\mathbb{R}^d \times \mathbb{R}^d} ||X - Y||^p \, \mathrm{d}\pi + \epsilon \mathrm{KL}(\pi|\mathbb{P} \otimes \mathbb{Q}) \\ &\text{with} \quad \mathrm{KL}(\pi|\mathbb{P} \otimes \mathbb{Q}) := \int_{\mathbb{R}^d \times \mathbb{R}^d} \log\left(\frac{\mathrm{d}\pi}{\mathrm{d}\mathbb{P}\mathrm{d}\mathbb{Q}}\right) \mathrm{d}\pi, \end{aligned} \tag{8}$$

with $\Gamma(\mathbb{P}, \mathbb{Q})$ the set of all joint probability measures whose marginals are $\mathbb{P}$, $\mathbb{Q}$ on $\mathbb{R}^d$, i.e. for all subsets $A \subset \mathbb{R}^d$, we have $\pi(A \times \mathbb{R}^d) = \mathbb{P}(A)$ and $\pi(\mathbb{R}^d \times A) = \mathbb{Q}(A)$. Here, ϵ mediates the strength of the Kullback-Leibler divergence.

The Sinkhorn divergence can be efficiently calculated for $\epsilon > 0$ (Cuturi, 2013). For $p = 2$ and $\epsilon = 0$ we can retrieve the quadratic Wasserstein distance and in the limit $\epsilon \to +\infty$ it becomes the MMD (Genevay *et al.*, 2017). BITES tunes ϵ to take advantage of the more stable OT gradients to improve the overlap while remaining computationally efficient. In the following, we denote it by $\mathrm{IPM}_\epsilon^p(\cdot, \cdot)$ to highlight the possibility to use any representation of the IPM. A thorough discussion of the Sinkhorn divergence, its theoretical properties, as well as one- and two-dimensional examples can be found in Feydy *et al.* (2018).

2.4 BITES

BITES model architecture: BITES combines survival modeling with counterfactual reasoning, i.e. it facilitates the development of treatment recommender systems using time-to-event data. BITES uses the network architecture shown in Figure 1 with loss function

$$\begin{aligned} &l_{\mathrm{BITES}}(\boldsymbol{x}_i, y_i, E_i, T_i) = \\ &\quad q\mathcal{L}_{\mathrm{Cox}}^{T=0}(h_0(\Phi^{T=0}(\mathbf{x})), Y^{T=0}, E^{T=0}) \\ &\quad +(1-q)\mathcal{L}_{\mathrm{Cox}}^{T=1}(h_1(\Phi^{T=1}(\mathbf{x})), Y^{T=1}, E^{T=1}) \\ &\quad +\alpha\mathcal{L}_{\mathrm{IPM}_\epsilon^p}(\Phi^{T=1}, \Phi^{T=0}), \end{aligned} \tag{9}$$

where q is the fraction of patients in the control cohort (patients with $T=0$) and $\mathcal{L}_{\mathrm{Cox}}^T$ is given by the negative Cox partial log-likelihood of Equation 2, where we parametrize the hazard function $h_T(\Phi(\mathbf{x}))$ according to the network architecture illustrated in Figure 1. The latent representation Φ is regularized by an IPM term to reduce differences between treatment and control distributions of non-confounding variables. Throughout the article, we used the Sinkhorn divergence of the smoothed OT loss with $p = 2$ as IPM term. Hence, the parameter ϵ in Equation 9 calibrates between the quadratic-Wasserstein distance ($\epsilon = 0$) and MMD ($\epsilon = \infty$). The total strength of the IPM regularization is adjusted by the hyperparameter α. Models with $\alpha = 0$ do not balance treatment effects and therefore we denote this method as 'Individual Treatment Effects for Survival' (ITES). Models with $\alpha > 0$ will be denoted as 'Balanced Individual Treatment Effects for Survival' (BITES). Large α values enforce balanced distributions between treatment and control population. Note, there is a trade-off between balancing distributions and model performance since outcome relevant information can be predictive for the treatment. (B)ITES uses the time-dependent ITE for treatment decisions. For the studies shown in this article, we assigned treatments based on the ITE evaluated for a survival probability of 50%, i.e. $\tau(\boldsymbol{x}_i) = (S(\boldsymbol{x})\lambda_1(y))^{-1}(0.5) - (S(\boldsymbol{x})\lambda_0(y))^{-1}(0.5)$.

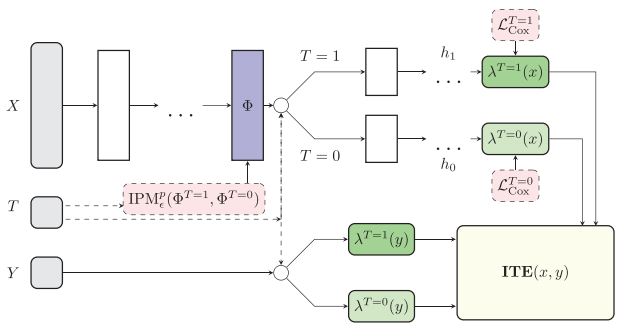

Fig. 1. The BITES network architecture. BITES uses shared deeply connected layers for both treatment options, which are mapped on a latent representation Φ. This is regularized by a Sinkhorn divergence to account for imbalances between treatment and control distributions. The factual and counterfactual proportional hazard rates are modeled by two different outcome heads (h_1 and h_0), respectively. These are used to predict the ITE together with the corresponding baseline hazard function. The latter is individually estimated for treatment and control patients

Implementation: BITES uses a deep architecture of dense-connected layers which are each followed by a dropout (Srivastava *et al.*, 2014) and a batch normalization layer (Ioffe and Szegedy, 2015). It uses ReLU activation functions (Nair and Hinton, 2010) and is trained using the Adam optimizer (Kingma and Ba, 2014). Further, early stopping based on non-decreasing validation loss and weight decay regularizations (Krogh and Hertz, 1992) are used to improve generalization. Our implementation is based on the *PyTorch* machine learning library (Paszke *et al.*, 2019) and the *pycox* package (Kvamme and Borgan, 2019). The Sinkhorn divergence is implemented using the *GeomLoss* package (Feydy *et al.*, 2018). We provide an easy-to-use python implementation which includes a hyperparameter optimization using the *ray[tune]* package (Liaw *et al.*, 2018) to efficiently distribute model training.

2.5 Treatment recommender systems

For comparison, we evaluated several strategies to build treatment recommender systems.

Cox regression model: We implemented the Cox regression as T-learner with treatment-specific survival models using *lifelines* (Davidson-Pilon *et al.*, 2021). Note, an ordinary Cox regression model which uses both the covariates $\mathcal{X}$ and the treatment variable $\mathcal{T}$ as predictor variables generally recommends the treatment with the better ATE; a treatment-specific term adds to $\boldsymbol{\beta}^T\boldsymbol{x}$ and thus the treatment which reduces the hazard most will be recommended. Therefore, we did not include the latter approach and focus on the Cox T-learner in our analysis.

DeepSurv: Katzman *et al.* (2018) suggested to provide individual recommendations based on single model predictions using $\mathcal{T}$ and $\mathcal{X}$ as covariates based on $\tau_{\text{DS}}(T, \boldsymbol{x}_i) = h_\theta(T = 1, \boldsymbol{x}_i) - h_\theta(T = 0, \boldsymbol{x}_i)$. Hence, it uses a treatment independent baseline hazard which could compromise the performance (Bellera *et al.*, 2010; Xue *et al.*, 2013).

Treatment-specific DeepSurv models: To account for treatment-specific differences of baseline hazard functions, we also estimated DeepSurv as a T-learner (T-DeepSurv), i.e. we learned models stratified for treatments. We then evaluated the time-dependent ITE based on the survival function $\tau_{\text{T-DS}}(\boldsymbol{x}_i, y) = S^{T=1}(\boldsymbol{x}_i, y) - S^{T=0}(\boldsymbol{x}_i, y)$.

Treatment-specific Random Survival Forests: Analogously to the previous approach, we learned treatment-specific RSF (Athey *et al.*, 2016; Ishwaran *et al.*, 2008) using the implementation of *scikit-survival* (Pölsterl, 2020) to estimate the time-dependent ITE.

SurvITE: Curth *et al.* (2021) suggested to learn discrete-time treatment-specific conditional hazard functions, which were estimated using an individual outcome head for each time interval. (We employed their python implementation available under https://github.com/chl8856/survITE.) We evaluated the time-dependent ITE to assign treatments, as for the latter two methods.

2.6 Performance measures

We used different measures to assess the performance of treatment recommendation systems. This comprises both measures for the quantification of prediction performance and of treatment assignment. Discriminative performance was assessed using a time-dependent extension of Harrell's C-index (Harrell, 1982) to account for differing baseline hazards, which evaluates

$$\Pr\Big(S(y_i|x_i) \;<\; S(y_i|x_j) \,|\, y_i \;<\; y_j \;\&\; E_i = 1\Big), \tag{10}$$

for all samples i and j at all event times $y_i^{E=1}$ (Antolini *et al.*, 2005). This reduces to Harrell's C-index for strictly ordered survival curves. To quantify the performance of treatment recommendations, we used the Precision in Estimation of Heterogeneous Effect (PEHE) score (Hill, 2011), which is defined as the difference in residuals between factual and counterfactual outcome:

$$\epsilon_{\text{PEHE}} = \frac{1}{N}\sum_{n=0}^{N}\Big([y_1(x_n) - y_0(x_n)] - [\hat{y}_1(x_n) - \hat{y}_0(x_n)]\Big)^2. \tag{11}$$

Note, the PEHE score can only be calculated if both the factual and counterfactual outcomes are known, which is usually only the case in simulation studies. Therefore, we restricted its application to the latter. There, we further quantified the proportion of correctly assigned 'best treatments'.

3 Results

3.1 Simulation studies

We performed three exemplary simulation studies. First, we simulated a scenario where covariates affect survival only linearly. Second, we simulated data with additional non-linear dependencies, and, finally, we performed a simulation where the treatment assignments were biased by the covariates.

Linear simulation study: In analogy to (Alaa and van der Schaar, 2017) and (Lee *et al.*, 2018), we simulated a 20-dimensional covariate vector $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2) \sim \mathcal{N}(0, \boldsymbol{I})$ consisting of two 10-dimensional vectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, with corresponding survival times given by

$$\begin{aligned} Y^{T=0}(\boldsymbol{x}) &\sim \exp\left(\left[\boldsymbol{\gamma}_1^T\boldsymbol{x}_1 + \boldsymbol{\gamma}_1^T\boldsymbol{x}_2\right]\right), \\ Y^{T=1}(\boldsymbol{x}) &\sim \exp\left(\left[\boldsymbol{\gamma}_2^T\boldsymbol{x}_1 + \boldsymbol{\gamma}_1^T\boldsymbol{x}_2\right]\right). \end{aligned} \tag{12}$$

We set the parameters $\boldsymbol{\gamma}_1 = (0.1, \ldots, 0.1)^T$ and $\boldsymbol{\gamma}_2 = (15, 35, 55, 75, 95, 115, 135, 155, 175, 195)^T \cdot 10^{-2}$. The first term in the exponent is treatment dependent while the second term affects survival under both treatments identically. This simulation gives an overall positive ATE in ~64% of the patients. Survival times exceeding 10 years were censored to resemble common censoring at the end of a study. Of the remaining samples, 50% were censored at a randomly drawn fraction $f_c \sim \mathcal{U}(0, 1)$ of the true unobserved survival time. Samples were assigned randomly to the treated, $T = 1$, and control group, $T = 0$, without treatment administration bias. Finally, we added an error $\epsilon \sim \mathcal{N}(0, 0.1 \cdot \boldsymbol{I})$ to all covariates. Detailed information about hyper-parameter selection is given in the Supplementary Section S1.

Figure 2a shows the distributions of Harrell's C-index evaluated on 1000 test samples for 50 consecutive simulation runs. We observed that across all investigated sample sizes (x-axis) the T-learner Cox regression showed superior performance, closely followed by ITES and BITES. These three methods performed equally well for the larger sample sizes $n = 1800$ and $n = 2400$. We further investigated the proportion of correctly assigned treatments, Figure 2a, and PEHE scores, Supplementary Figure S1, where we obtained qualitatively similar trends. RSF, DeepSurv and T-DeepSurv showed inferior performance with respect to C-Indices, correctly assigned treatments, and PEHE scores. Results for lower sample sizes can be found in Supplementary Section S2. Consistent with previous findings, Cox regression outperformed all competitors closely followed by (B)ITES.

Non-linear simulation study: Next, we simulated non-linear treatment-outcome dependencies using the model

$$\begin{aligned} Y^{T=0}(\boldsymbol{x}) &\sim \exp\left(\left[(\boldsymbol{\gamma}_1^T\boldsymbol{x}_1)^2 + \boldsymbol{\gamma}_1^T\boldsymbol{x}_2\right]c\right), \\ Y^{T=1}(\boldsymbol{x}) &\sim \exp\left(\left[(\boldsymbol{\gamma}_2^T\boldsymbol{x}_1)^2 + \boldsymbol{\gamma}_1^T\boldsymbol{x}_2\right]c\right), \end{aligned} \tag{13}$$

where we set the parameters $\boldsymbol{\gamma}_1 = (2, \ldots, 2)^T$ and $\boldsymbol{\gamma}_2 = (0.5, 0.9, 1.3, 1.7, 2.1, 2.5, 2.9, 3.3, 3.7, 4.1)^T$. Note, the first term imposes sizable non-linear effects which differ between both treatments. We further scaled the polynomials by $c = 0.01$ to yield realistic survival times up to 10 years. This setting gives an overall positive ATE in ~64% of the patients.

Figure 2b gives the performance of the evaluated methods in terms of Harrell's C-index. We observed that the ordinary Cox regression with linear predictor variables performs worst across all sample sizes, followed by RSF, and SurvITE. Approximately equal performance was observed for the DeepSurv approaches, ITES, and BITES. Among these methods, the treatment-specific DeepSurv models (T-DeepSurv) showed a higher variance across the simulation runs, in particular for the low sample sizes. Next, we studied the corresponding PEHE scores (Supplementary Fig. S1) and the

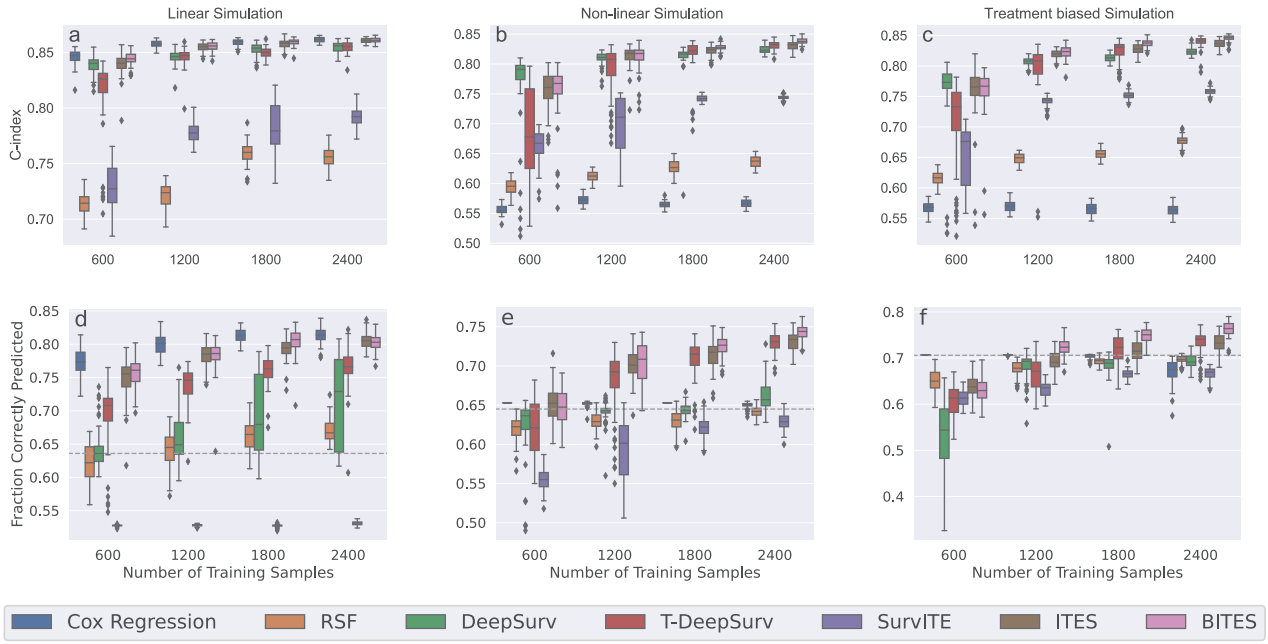

Fig. 2. Harrell's C-index and the fraction of correctly predicted treatments for the linear (**a, d**), non-linear (**b, e**), and treatment biased non-linear (**c, f**) simulations. The boxplots give the distribution for 50 consecutive simulation runs, i.e. for different model initializations, based on the best set of hyper-parameter determined by the validation C-index. Results are shown for different training sample sizes with 1000 fixed test samples for each of the simulations. The dashed horizontal line represents the fraction of patients that benefits for 100% treatment administration

proportion of correctly assigned treatments (Fig. 2e). We observed, although DeepSurv performed well in terms of C-Indices, that the performance was highly compromised in the latter two measures. In fact, it was not able to outperform the recommendation based on the ATE, i.e. always assigning $T=1$, which corresponds to the dashed horizontal line. We further observed that SurvITE performed worst in this scenario with both substantially lower proportions of correctly assigned treatments and higher PEHE scores compared to the other methods. Here, T-DeepSurv, ITES, and BITES performed best, however, the results of the former are inferior compared to ITES and BITES for sample sizes of $n=600$ and $n=1200$. Additional simulations for smaller sample sizes can be found in Supplementary Section S2, where we observed that none of the methods is able to outperform the ATE-based recommendation.

Non-linear simulation study with treatment bias: Finally, we repeated the non-linear simulation study but now took into account a treatment assignment bias, i.e. the value of one or more covariates is indicative of the applied treatment. To simulate this effect, we assigned the treatment with a 90% probability if the fifths entry of $\boldsymbol{x}_1$ or $\boldsymbol{x}_2$ was larger than zero. To ensure that the unconfoundedness assumption holds, we set the corresponding entries γ_1 and γ_2 to zero. This simulation study yields a positive treatment effect in ~71% of the patients (dashed horizontal line in Fig. 2f).

Figure 2c and f and Supplementary Figure S1 show the results in terms of C-index, correctly assigned treatments, and PEHE scores, respectively. Similar to the previous studies, the best performing methods with respect to C-Indices were the two DeepSurv models, ITES and BITES. With respect to correctly assigned treatments and PEHE scores, however, BITES consistently outperformed the other methods for reasonable sample sizes starting from $n=1200$. For $n=600$, none of the methods was able to outperform a model where the treatment is always recommended (dashed line in Fig. 2f). This was further confirmed for lower sample sizes (Supplementary Section S2).

3.2 Bites optimizes hormone treatment in patients with breast cancer

We retrieved data of 1,545 node-positive breast cancer patients from the Rotterdam Tumour Bank (Foekens *et al.*, 2000) as provided by Katzman *et al.* (2018). The latter data were preprocessed according to (Royston and Parmar, 2013). We used recurrence-free survival (RFS) time, defined as the time from primary surgery to the earlier of disease recurrence or death from any cause, as outcome for the further analysis. The available patient characteristics are age, menopausal status (pre/post), number of cancerous lymph nodes, tumor grade and progesterone and estrogen receptor status. Of these patients, 339 were treated by a combination of chemotherapy and hormone therapy. The remaining patients were treated by chemotherapy only. Note, in this study, the application of hormone treatment was not randomized. In total ~37% of the patients were censored.

Table 1. Predictive outcomes on the controlled randomized test set of the RGBSG data obtained by each of the discussed models with minimum validation loss found in a hyper-parameter grid search

Method	C-index	*P*-value	Fraction $T=1$
Cox reg.	0.471	0.0034	100%
DeepSurv	0.671	0.0034	100%
T-DeepSurv	0.652	0.2023	92.9%
RSF	0.675	0.0013	82.5%
SurvITE	0.631	0.0039	98.1%
ITES	**0.676**	0.000198	75.8%
BITES	0.666	**0.000016**	83.4%

Values in boldface indicate the best performing model with respect to C-index and *P*-value, respectively.

We used these data to learn treatment recommender systems in order to predict the ITE of adding hormone therapy to chemotherapy. We performed hyper-parameter tuning as outlined in Supplementary Section S3, and selected the models with the lowest validation loss, respectively.

Next, we evaluated the performance using test data from the GBSG Trial 2 (Schmoor *et al.*, 1996). Excluding cases with missing covariates, it contains 686 individual patients, with ~65% randomized hormone treatment assignments. The obtained C-indices are summarized in Table 1. Note, since only the factual outcomes are observable, we could not evaluate the performance with respect to correctly assigned 'best treatments' or PEHE scores. However, to

substantiate our findings, we stratified our patients into two groups; the group 'recommended treatment' contains samples where the recommended treatment coincides with the applied treatment, while the group 'anti-recommended treatment' contains the samples where the recommended treatment does not coincide with the applied treatment (following Katzman *et al.*, 2018). The corresponding Kaplan–Meier (KM) curves of BITES are shown in Figure 3 with recommended treatment in green and anti-recommended treatment in red. Corresponding results for the other methods are shown in Supplementary Figure S3. For comparison, KM curves for the treated and control group are shown in blue and orange in Figure 3. Interestingly, BITES recommends hormone treatment only in 83.4% which resulted in the largest difference in survival based on the recommendations made by BITES ($P = 0.000016$). On the other hand, DeepSurv and Cox regression suggest to treat all patients with hormone therapy, closely followed by SurvITE (treatment recommended for 98.1% of patients). The results for all models are summarized in Table 1. Note, the group with BITES recommendation showed a superior survival compared to the treated group and the group with BITES anti-recommendation showed an inferior performance compared to the control group. Both comparisons, however, were not significant in a log-rank test.

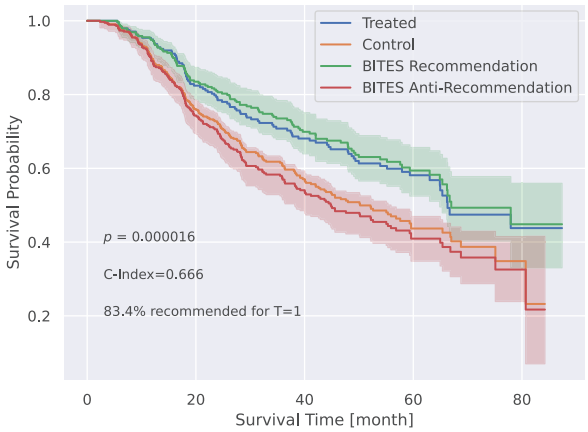

Fig. 3. Recurrence-free survival probability for patients grouped according to the respective treatment recommendations of BITES, based on the test data from the GBSG Trial 2. For comparison, we show the KM curves for all hormone treated and untreated (control) patients in blue and orange, respectively (shown without error bars for better visibility)

Finally, we explored feature importance of the BITES model using SHAP values (Lundberg and Lee, 2017) with results shown in Figure 4 which correspond to treatment option $T = 0$ (no hormone treatment) and $T = 1$ (hormone treatment), respectively. Here, points correspond to patients and positive (negative) SHAP values on the x-axis indicate an increased (decreased) risk of failure. Further, the feature value is illustrated in colors ranging from red to blue, where high values are shown in red and low values in blue. We observed that the number of positive lymph nodes has the strongest impact on survival with SHAP values ranging from ~ -0.5 to ~ 1 in the group with and without hormone treatment, where more positive lymph nodes (shown in red) indicate a worse survival. Considering menopausal status, we observed an increased risk of death and recurrence in postmenopausal breast cancer patients that had not received adjuvant hormone treatment ($T = 0$). This effect was substantially mitigated in the hormone-treated group, which is in line with the observation that postmenopausal, more than premenopausal breast cancer patients draw a disease-free survival benefit from extended adjuvant endocrine treatment (Li *et al.*, 2018). However, this does not preclude a survival benefit from hormone treatment for certain premenopausal breast cancer patients as revealed by a comparison of Figure 4a and b. It is also noteworthy, that high tumor grade (grade 3, shown in red) yielded increased SHAP values of up to 0.5 in the tamoxifen-treated group. This effect was substantially mitigated in the group without hormone treatment. This finding is in line with a recent study by Dar *et al.* (2021), which found a significant tamoxifen treatment benefit only among patients suffering from lower grade tumors, while no benefit was observed for grade 3 tumors. In summary, we observed strong hints that hormone treatment alleviates the negative effect of menopause, and increases the negative effect of high tumor grade on patient survival.

4 Conclusion

We presented BITES, which is a machine learning framework to optimize individual treatment decisions based on time-to-event data. It combines Deep Neural Network counterfactual reasoning with Cox's PH model. It further enables balancing of treated and non-treated patients using IPM on a latent layer data representation. We demonstrated in simulation studies that BITES outcompetes state-of-the-art methods with respect to prediction performance (Harrell's C-index), correctly assigned treatments, and PEHE scores. We

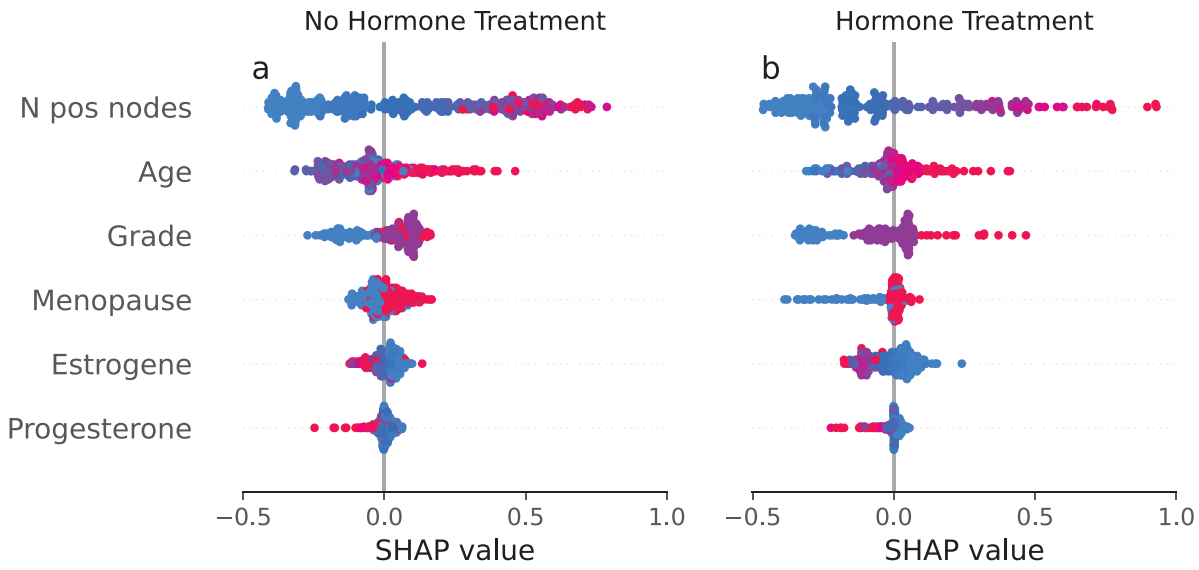

Fig. 4. SHAP (SHapley Additive exPlanations) values for the best selected BITES model on the controlled randomized test samples of the RGBSG data. Red points correspond to high and blue points to low feature values. A positive SHAP value indicates an increased hazard and hence decreased survival chances and vice versa (A color version of this figure appears in the online version of this article)

observed that BITES can effectively capture both linear and non-linear covariate outcome dependencies on both small and large scale observational studies. Moreover, we showed that BITES can be used to optimize hormone treatment in breast cancer patients. Using independent data from the GBSG Trial 2, we observed that BITES treatment recommendations might improve patients' RFS. In this context, SHAP values were demonstrated to enhance the interpretability and transparency of treatment recommendations.

Like most recently developed counterfactual tools, BITES depends on the *strong ignorability* assumption. Hence, caution is necessary when analyzing heavily confounded observational data. Future work needs to address more specialized time-to-event models, such as competing event models, and the generalization to multiple treatments and combinations thereof. Both could substantially broaden the scope of applications for BITES.

In summary, BITES facilitates treatment optimization from time-to-event data. In combination with SHAP values, BITES models can be easily interpreted on the level of individual patients, making them a versatile backbone for treatment recommender systems.

Funding

This work was supported by the German Federal Ministry of Education and Research (BMBF) within the framework of the e:Med research and funding concept (grants 01ZX1912A, 01ZX1912C).

Conflict of Interest: none declared.

References

Alaa,A.M. and van der Schaar,M. (2017). Bayesian inference of individualized treatment effects using multi-task Gaussian processes. In: *31st Conference on Neural Information Processing Systems, Long Beach, California, USA*.

Antolini,L. *et al.* (2005) A time-dependent discrimination index for survival data. *Stat. Med.*, **24**, 3927–3944.

Atan,O. *et al.* (2018). Deep-treat: learning optimal personalized treatments from observational data using neural networks. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, p. 32.

Athey,S. and Wager,S. (2019) Estimating treatment effects with causal forests: an application. *Obs. Stud.*, **5**, 37–51. https://doi.org/10.1353/obs.2019.0001.

Athey,S. *et al.* (2016). Generalized random forests. *Ann. Stat.* **47**, 1148–1178.

Bellera,C.A. *et al.* (2010) Variables with time-varying effects and the Cox model: some statistical concepts illustrated with a prognostic factor study in breast cancer. *BMC Med. Res. Methodol.*, **10**, 20.

Bottou,L. *et al.* (2013) Counterfactual reasoning and learning systems: the example of computational advertising. *J. Mach. Learn. Res.*, **14**, 3207–3260.

Breslow,N. (1972) Discussion on Professor Cox's paper. *J. R. Stat. Soc. Series B (Methodol.)*, **52**, 216–217.

Chapfuwa,P. *et al.* (2020). Enabling counterfactual survival analysis with balanced representations. In: *Proceedings of the Conference on Health, Inference, and Learning*, pp. 133–145.

Cox,D.R. (1972) Regression models and Life-Tables. *J. R. Stat. Soc. Series B (Methodol.)*, **34**, 187–220.

Curth,A. *et al.* (2021). SurvITE: learning heterogeneous treatment effects from time-to-event data. *Adv. Neural Inf. Process. Syst.*, **34**.

Cuturi,M. (2013) Sinkhorn distances: lightspeed computation of optimal transport. *Adv. Neural Inf. Process. Syst.*, **26**.

D'Amour,A. *et al.* (2017). Overlap in observational studies with high-dimensional covariates. *J. Econometrics*, **221**, 644–654.

Dar,H. *et al.* (2021) Assessment of 25-year survival of women with estrogen receptor-positive/erbb2-negative breast cancer treated with and without tamoxifen therapy: a secondary analysis of data from the Stockholm tamoxifen randomized clinical trial. *JAMA Netw. Open.*, **4**, e2114904.

Davidson-Pilon,C. *et al.* (2021). *CamDavidsonPilon/lifelines: 0.26.0*. Zenodo.

Diamond,A. and Sekhon,J.S. (2013) Genetic matching for estimating causal effects: a general multivariate matching method for achieving balance in observational studies. *Rev. Econ. Stat.*, **95**, 932–945.

Faraggi,D. and Simon,R. (1995) A neural network model for survival data. *Stat. Med.*, **14**, 73–82.

Feydy,J. *et al.* (2018) Interpolating between Optimal Transport and MMD using Sinkhorn Divergences. In: *The 22nd International Conference on Artificial Intelligence and Statistics*, Vol. 22, pp. 2681–2690.

Foekens,J.A. *et al.* (2000) The urokinase system of plasminogen activation and prognosis in 2780 breast cancer patients. *Cancer Res.*, **60**, 636–643.

Frieden,T.R. (2017) Evidence for health decision making—beyond randomized, controlled trials. *N Engl. J. Med.*, **377**, 465–475.

Genevay,A. *et al.* (2017) Learning generative models with Sinkhorn divergences. In: *International Conference on Artificial Intelligence and Statistics*, pp. 1608–1617.

Gensheimer,M.F. and Narasimhan,B. (2019) A scalable discrete-time survival model for neural networks. *PeerJ*, **7**, e6257.

Gretton,A. *et al.* (2012) A kernel two-sample test. *J. Mach. Learn. Res.*, **13**, 723–773.

Harrell,F.E. (1982) Evaluating the yield of medical tests. *JAMA*, **247**, 2543.

Hill,J.L. (2011) Bayesian nonparametric modeling for causal inference. *J. Comput. Graph. Stat.*, **20**, 217–240.

Holland,P.W. (1986) Statistics and causal inference. *J. Am. Stat. Assoc.*, **81**, 945–960.

Huang,C. *et al.* (2016) Local similarity-aware deep feature embedding. *Adv. Neural Inf. Process. Syst.*, **29**.

Imai,K. and Ratkovic,M. (2014) Covariate balancing propensity score. *J. R Stat. Soc. B.*, **76**, 243–263.

Ioffe,S. and Szegedy,C. (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift. In: *International conference on machine learning*, pp. 448–456.

Ishwaran,H. *et al.* (2008) Random survival forests. *Ann. Appl. Stat.*, **2**, 841–860.

Jachimowicz,R.D. *et al.* (2021) Gene expression-based outcome prediction in advanced stage classical hodgkin lymphoma treated with BEACOPP. *Leukemia*, **35**, 3589–3593.

Johansson,F.D. *et al.* (2016) Learning representations for counterfactual inference. In: *International conference on machine learning*, pp. 3020–3029.

Johansson,F.D. *et al.* (2020) Generalization bounds and representation learning for estimation of potential outcomes and causal effects. arXiv preprint arXiv:2001.07426.

Kantorovitch,L. (1958) On the translocation of masses. *Manage. Sci.*, **5**, 1–4.

Katzman,J. *et al.* (2018) DeepSurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC Med. Res. Methodol.*, **18**, 1.

King,G. and Nielsen,R. (2019) Why propensity scores should not be used for matching. *Polit. Anal.*, **27**, 435–454.

Kingma,D.P. and Ba,J. (2014) Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kohavi,R. *et al.* (2009) Controlled experiments on the web: survey and practical guide. *Data Min. Knowl. Disc.*, **18**, 140–181.

Krogh,A. and Hertz,J. (1992) A simple weight decay can improve generalization. In: Moody,J., Hanson,S. and Lippmann, R.P. (eds.) *Advances in Neural Information Processing Systems*, Vol. 4. Morgan-Kaufmann, San Mateo CA, United States, pp. 950–957.

Kvamme,H. and Borgan,Ø. (2019) Continuous and discrete-time survival prediction with neural networks. *Lifetime Data Anal.*, **27**, 710–736.

Lee,C. *et al.* (2018) DeepHit: a deep learning approach to survival analysis with competing risks. In: *Proceedings of the AAAI conference on artificial intelligence*, p. 32.

Li,L. *et al.* (2018) Clinical outcomes comparison of 10 years versus 5 years of adjuvant endocrine therapy in patients with early breast cancer. *BMC Cancer*, **18**, 977.

Li,S. and Fu,Y. (2017) Matching on balanced nonlinear representations for treatment effects estimation. *Adv. Neural Inf. Process. Syst.*, **30**.

Liaw,R. *et al.* (2018) Tune: a research platform for distributed model selection and training. arXiv preprint arXiv:1807.05118.

Lu,D. *et al.* (2020) Reconsidering generative objectives for counterfactual reasoning. *Adv. Neural Inf. Process. Syst.*, **33**, 21539–21553.

Lundberg,S. and Lee,S.-I. (2017) A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.*, **32**.

Martinussen,T. and Scheike,T.H. (2006). *Dynamic Regression Models for Survival Data*. Springer, New York, NY.

Müller,A. (1991) Integral probability metrics and their enerating classes of functions. *Adv. Appl. Probab.*, **29**, 429–443.

Nair,V. and Hinton,G.E. (2010) Rectified linear units improve restricted Boltzmann machines. In: *International Conference on Machine Learning*.

Paszke,A. *et al.* (2019) PyTorch: an imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.*, **32**.

Pearl,J. (2008) *Causality: Models, Reasoning and Interference*, 8. pr edition. Cambridge University Press, New York.

Pearl,J. (2009) Causal inference in statistics: an overview. *Statist. Surv.*, **3**, 96–146.

Pearl,J. (2017) Detecting latent heterogeneity. *Sociol. Methods Res.*, **46**, 370–389.

Pölsterl,S. (2020) Scikit-survival: a library for time-to-Event analysis built on top of scikit-learn. *J. Mach. Learn. Res.*, **21**, 1–6.

Ramdas,A. *et al.* (2017) On Wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, **19**, 47.

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

Royston,P. and Parmar,M.K.B. (2013) Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC Med. Res. Methodol.*, **13**, 152.

Rubin,D.B. (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.*, **66**, 688–701.

Schmoor,C. *et al.* (1996) Randomized and non-randomized patients in clinical trials: experiences with comprehensive cohort studies. *Stat. Med.*, **15**, 263–271.

Schumacher,M. *et al.* (1994) Randomized 2 x 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. German breast cancer study group. *JCO*, **12**, 2086–2093.

Shalit,U. *et al.* (2016) Estimating individual treatment effect: generalization bounds and algorithms. In: I*nternational Conference on Machine Learning*, pp. 3076–3085.

Sriperumbudur,B.K. *et al.* (2012) On the empirical estimation of integral probability metrics. *Electron. J. Stat.*, **6, 1550–1599.**

Srivastava,N. *et al.* (2014) Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, **15**, 1929–1958.

Staiger,A.M. *et al.*; German High Grade Non-Hodgkin's Lymphoma Study Group (DSHNHL) (2020) A novel lymphoma-associated macrophage interaction signature (LAMIS) provides robust risk prognostication in diffuse large B-cell lymphoma clinical trial cohorts of the DSHNHL. *Leukemia*, **34**, 543–552.

Tibshirani,R. (1997) The lasso method for variable selection in the cox model. *Stat. Med.*, **16**, 385–395.

Wager,S. and Athey,S. (2017) Estimation and inference of heterogeneous treatment effects using random forests. *J Am Stat Assoc*, **113**, 1228–1242.

Xue,X. *et al.* (2013) Testing the proportional hazards assumption in case-cohort analysis. *BMC Med. Res. Methodol.*, **13**, 88.

Yao,L. *et al.* (2018) Representation learning for treatment effect estimation from observational data. *Adv. Neural Inf. Process. Syst.*, **31**.

Yoon,J. *et al.* (2018) GANITE: estimation of individualized treatment effects using generative adversarial nets. In: *International Conference on Learning Representations, Vancouver, BC, Canada.*

Zacharias,H.U. *et al.* (2021) A predictive model for progression of CKD to kidney failure based on routine laboratory tests. *Am. J. Kidney Dis.*, **79**, 217–230.e1.