BITES: Balanced Individual Treatment Effect for Survival data

Stefan Schrod [1]   Andreas Schäfer [2]   Stefan Solbrig [2]   Robert Lohmayer [2]   Wolfram Gronwald [3]   Peter J. Oefner [3]
Tim Beißbarth [1]   Rainer Spang [4]   Helena U. Zacharias [5][6]   Michael Altenbuchinger [1]

Abstract

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 is 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). We show in simulation studies that this approach outperforms the state of the art. Further, 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.

BITES is provided as an easy-to-use python implementation.

1. Introduction

Inferring the effect of interventions on outcomes is relevant in diverse domains, comprising precision medicine and epidemiology (Frieden, 2017), or marketing (Kohavi et al., 2009; Bottou et al., 2013). A fundamental issue of causal reasoning is that potential outcomes are observed only for the applied intervention but not for its alternatives (the counterfactuals). For instance, in medicine, only the factual treatment outcome is observed. The counterfactual outcomes remain hidden.

Estimates of Average Treatment Effects (ATE) do not necessarily hold on the level of individual patients, and the Individual Treatment Effect (ITE) 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; 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; Wager and Athey, 2017; Athey and Wager, 2019) 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 towards treatment-specific distributions. This issue was recently approached by several groups which balanced the treated and non-treated distributions

[1]Department of Medical Bioinformatics, University Medical Center Göttingen, Göttingen, Germany [2]Institute of Theoretical Physics, University of Regensburg, Regensburg, Germany [3]Institute of Functional Genomics, University of Regensburg, Regensburg, Germany [4]Department of Statistical Bioinformatics, Institute of Functional Genomics, University of Regensburg, Regensburg, Germany [5]Institute of Clinical Molecular Biology, Kiel University and University Medical Center Schleswig-Holstein, Campus Kiel, Kiel, Germany [6]Department of Internal Medicine I, University Medical Center Schleswig-Holstein, Campus Kiel, Kiel, Germany. Correspondence to: Stefan Schrod <stefan.schrod@bioinf.med.uni-goettingen.de>, Michael Altenbuchinger <michael.altenbuchinger@bioinf.med.uni-goettingen.de>.

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 autoencoder 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 (Cox, 1972; Breslow, 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 by Tibshirani, (1997), with applications ranging from the prediction of adverse events by 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 (Lee et al., 2018; Gensheimer and Narasimhan, 2019; Kvamme and Borgan, 2019) and random survival forests (RSF) (Ishwaran et al., 2008; Athey and Wager, 2019).

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 out-

perform 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[1].

2. 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 proportional hazards 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. Further, 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

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

and corresponds to the risk of dying at time y (Cox, 1972), i.e. a greater hazard corresponds to greater risk of failure.

---

[1] https://github.com/sschrod/BITES

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 proportional hazards (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 (Cox, 1972; Breslow, 1972) is given by:

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

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 $\mathcal{Y} = \mathcal{Y}_0 \times \ldots \times \mathcal{Y}_{(k-1)}$. 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

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

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

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

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

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).

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

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

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 (Johansson et al., 2016; Shalit et al., 2016; Li and Fu, 2017; Yao et al., 2018; Huang et al., 2016; Johansson et al., 2020; Lu et al., 2020; D'Amour et al., 2017). 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)

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

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,

$$S_\epsilon^p(\mathbb{P}, \mathbb{Q}) := W_\epsilon^p(\mathbb{P}, \mathbb{Q}) - \frac{1}{2} W_\epsilon^p(\mathbb{P}, \mathbb{P}) - \frac{1}{2} W_\epsilon^p(\mathbb{Q}, \mathbb{Q}),$$

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

Definition 2.1 (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*

$$W_\epsilon^p(\mathbb{P}, \mathbb{Q}) := \min_{\pi \in \Gamma(\mathbb{P}, \mathbb{Q})} \int_{\mathbb{R}^d \times \mathbb{R}^d} ||X - Y||^p \; d\pi$$
$$+ \epsilon KL(\pi | \mathbb{P} \otimes \mathbb{Q})$$
$$with \quad KL(\pi | \mathbb{P} \otimes \mathbb{Q}) := \int_{\mathbb{R}^d \times \mathbb{R}^d} \log \left(\frac{d\pi}{d\mathbb{P} \; d\mathbb{Q}} \right) \; d\pi,$$

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 $\epsilon \rightarrow +\infty$ 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 $\text{IPM}_\epsilon^p(\cdot, \cdot)$ to highlight the possibility to use any representation of the IPM. A thorough discussion of Sinkhorn divergences with 1- and 2-dimensional examples can be found in (Feydy et al., 2018).

2.4. 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 individual treatment effect based
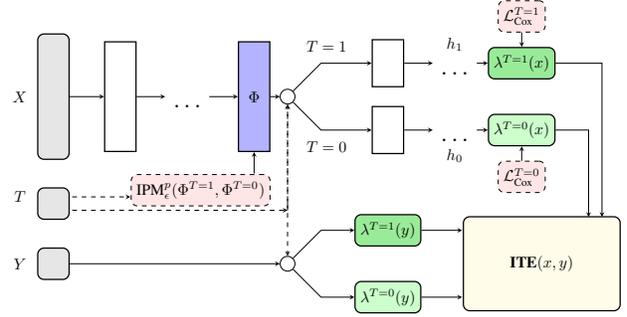

Figure 1. The BITES network architecture.

on the survival functions $\tau_{\text{T-DS}}(\boldsymbol{x}_i, y) = S^{T=1}(\boldsymbol{x}_i, y) - S^{T=0}(\boldsymbol{x}_i, y)$.

Treatment-specific Random Survival Forests Analogously to the previous approach, we learned treatment-specific Random Survival Forests (RSF) (Ishwaran et al., 2008; Athey et al., 2016) 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[2]. We evaluated the time-dependent ITE to assign treatments, as for the latter two methods.

2.5. 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

$$l_{\text{BITES}}(\boldsymbol{x}_i, y_i, E_i, T_i) =$$
$$q \mathcal{L}_{\text{Cox}}^{T=0}(h_0(\Phi(\mathbf{x})), Y^{T=0}, E^{T=0})$$
$$+ (1 - q) \mathcal{L}_{\text{Cox}}^{T=1}(h_1(\Phi(\mathbf{x})), Y^{T=1}, E^{T=1})$$
$$+ \alpha \mathcal{L}_{\text{IPM}_\epsilon^p}(\Phi^{T=1}, \Phi^{T=0}), \quad (2)$$

where q is the fraction of patients in the control cohort (patients with $T = 0$) and $\mathcal{L}_{\text{Cox}}^T$ is given by the negative Cox partial log-likelihood of Equation 1, 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

---

[2]We employed their python implementation available under `https://github.com/chl8856/survITE`.

Sinkhorn divergence of the smoothed OT loss with $p = 2$ as IPM term. Hence, the parameter ϵ in Equation 2 calibrates between the quadratic-Wasserstein distance ($\epsilon = 0$) and MMD ($\epsilon = \infty$). The total strength of the IPM regularization is adjusted by 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). (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. $\tau(\boldsymbol{x_i}) = (S(\boldsymbol{x})\lambda_1(y))^{-1}(0.5) - (S(\boldsymbol{x})\lambda_0(y))^{-1}(0.5)$.

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.

2.6. 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

$$\Pr\left(S(y_i|x_i) < S(y_i|x_j) \mid y_i < y_j \;\&\; E_i = 1\right), \quad (3)$$

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 Heterogenous Effect (PEHE) score (Hill, 2011), which is defined as the difference in residuals between factual and counterfactual outcome:

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

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 et al., (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

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

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 average treatment effect in $\sim 64\%$ 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 $\boldsymbol{\epsilon} \sim \mathcal{N}(0, 0.1 \cdot \boldsymbol{I})$ to all covariates. Detailed information about hyper-parameter selection is given in the Supplementary materials.

Figure 2(a) 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 2(a), 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.
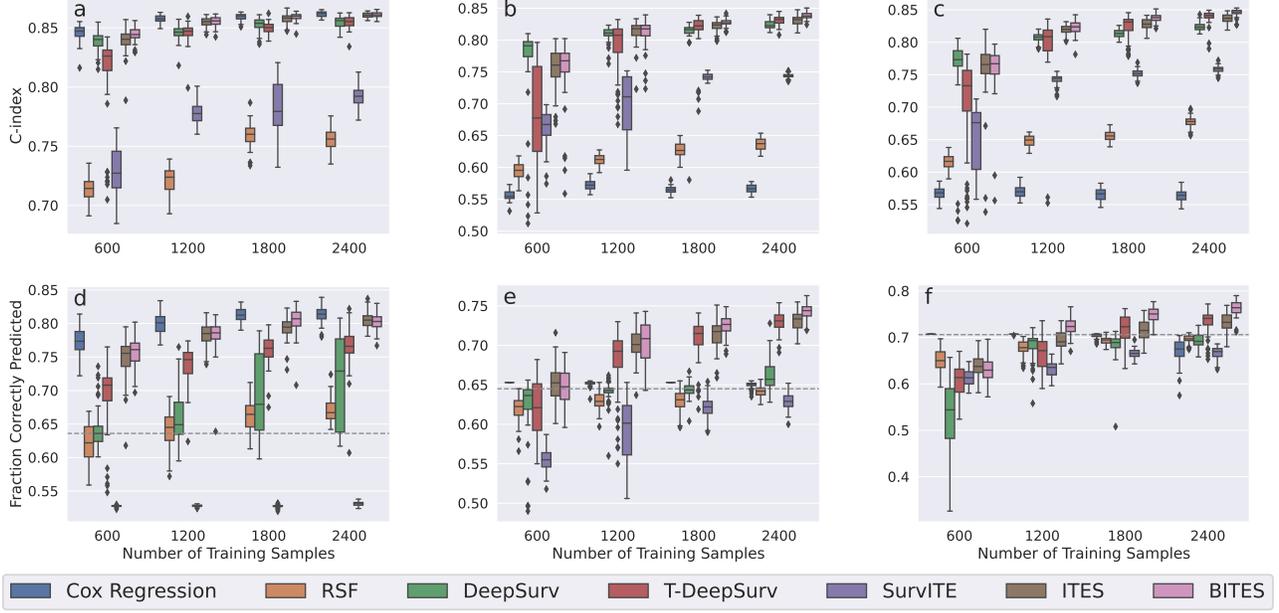
Figure 2. Harrell's C-index and the fraction of correctly predicted treatments for the linear (a,d), non-linear (b,e), and treatment biased non-linear (c,f) simulations. The boxplots give the distribution for 50 consecutive simulation runs, i.e. for different model initializations, based on the best 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.

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

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

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 average treatment effect in $\sim 64\%$ of the patients.

Figure 2(b) 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 Figure S1) and the proportion of correctly assigned treatments (Figure 2(e)). 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 $n = 1200$.

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 $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ to zero. This simulation study yields a positive treatment effect in $\sim 71\%$ of the patients (dashed horizontal line in Figure 2(f)).

Figures 2(c), 2(f), and Supplementary Figure S1, show the results in terms of C-index, correctly assigned treatments,
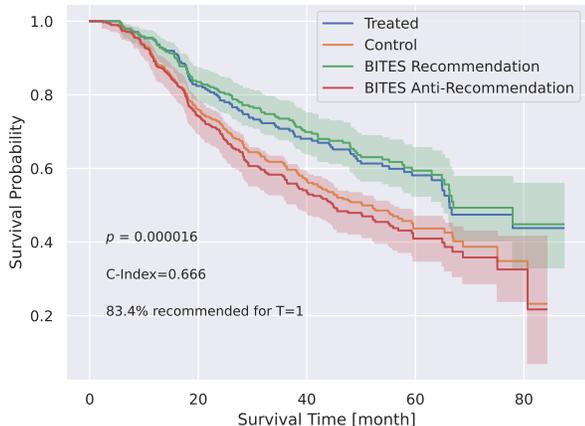
Figure 3. Survival probability for patients grouped according to the respective treatment recommendations of BITES, based on the test data from the GBSG Trial 2. For comparison, we show the KM curves for all hormone treated and untreated (control) patients in blue and orange, respectively (shown without error bars for better visibility).

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	0.202	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%

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 Figure 2(f)).

3.2. BITES optimizes hormone treatment in patients with breast cancer

We retrieved pre-processed data of 1,545 breast cancer patients as used by Katzman et al., (2018), which were originally extracted from the Rotterdam Tumour Bank (Foekens et al., 2000). 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 $\sim 37\%$ of the patients were censored.

We used these data to learn treatment recommender systems in order to predict the individual treatment effect of adding hormone therapy to chemotherapy. We performed hyper-parameter tuning as outlined in the Supplementary Material, 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 $\sim 65\%$ 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 S2. 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.

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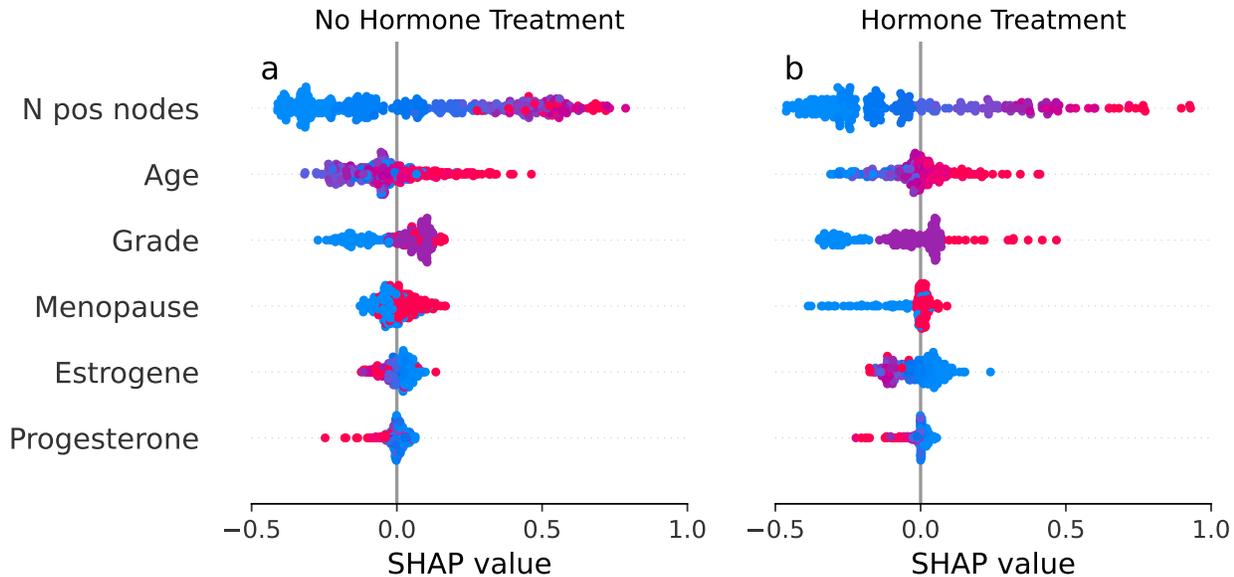
Figure 4. SHAP (SHapley Additive exPlanations) values for the best selected BITES model on the controlled randomized test samples of the RGBSG data. Red points correspond to high and blue points to low feature values. A positive SHAP value indicates an increased hazard and hence decreased survival chances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

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 the menopausal status, we observed that patients post-menopause showed increased risk of death in the group without hormone treatment ($T = 0$). Interestingly, this effect was substantially mitigated in the hormone-treated group. It was particularly interesting to observe that high tumor grade (grade 3, shown in red), yields high SHAP values up to 0.5 in the hormone treated group. This effect was substantially mitigated in the group without hormone treatment. 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 proportional hazards model. It further enables balancing of treated and non-treated patients using integral probability metrics 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), cor-

rectly assigned treatments, and PEHE scores. We 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 patient survival. 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.

References

Ahmed M Alaa and Mihaela van der Schaar. Bayesian Inference of Individualized Treatment Effects using Multi-task Gaussian Processes. *31st Conference on Neural Information Processing Systems*, 2017.

Laura Antolini, Patrizia Boracchi, and Elia Biganzoli. A time-dependent discrimination index for survival data. *Statistics in medicine*, (24):3927–3944, 2005.

Onur Atan, James James Jordon, and Mihaela va der Schaar. Deep-Treat: Learning Optimal Personalized Treatments From Observational Data Using Neural Networks. 2018.

Susan Athey and Stefan Wager. Estimating Treatment Effects with Causal Forests: An Application, 2019. URL http://arxiv.org/pdf/1902.07409v1.

Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized Random Forests, 2016. URL https://arxiv.org/pdf/1610.01271.

Carine A. Bellera, Gaëtan MacGrogan, Marc Debled, Christine Tunon de Lara, Véronique Brouste, and Simone Mathoulin-Pélissier. Variables with time-varying effects and the Cox model: some statistical concepts illustrated with a prognostic factor study in breast cancer. *BMC Medical Research Methodology*, 10(1):20, 2010.

Léon Bottou, Jonas Peters, and Joaquin Quinonero-Candela. Counterfactual Reasoning and Learning Systems: The Example of Computational Advertising. *Journal of Machine Learning Research 14*, pages 3207–3260, 2013.

Norman Breslow. Discussion on Professor Cox's Paper. *Journal of the Royal Statistical Society: Series B (Methodological)*, (34):216–217, 1972.

Paidamoyo Chapfuwa, Serge Assaad, Shuxi Zeng, Michael J. Pencina, Lawrence Carin, and Ricardo Henao. Enabling Counterfactual Survival Analysis with Balanced Representations, 2020. URL https://arxiv.org/pdf/2006.07756.

David R Cox. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*,, (Vol 34, No 2):187–220, 1972.

Alicia Curth, Changhee Lee, and Mihaela van der Schaar. SurvITE: Learning Heterogeneous Treatment Effects from Time-to-Event Data, 2021. URL https://arxiv.org/pdf/2110.14001.

Marco Cuturi. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. *Neural Information Processing Systems*, 2013.

Alexander D'Amour, Peng Ding, Avi Feller, Lihua Lei, and Jasjeet Sekhon. Overlap in Observational Studies with High-Dimensional Covariates, 2017. URL https://arxiv.org/pdf/1711.02582.pdf.

Cameron Davidson-Pilon, Jonas Kalderstam, Noah Jacobson, Sean Reed, Ben Kuhn, Paul Zivich, Mike Williamson, AbdealiJK, Deepyaman Datta, Andrew Fiore-Gartland, Alex Parij, Daniel WIlson, Gabriel, Luis Moneda, Arturo Moncada-Torres, Kyle Stark, Harsh Gadgil, Jona, JoseLlanes, Karthikeyan Singaravelan, Lilian Besson, Miguel Sancho Peña, Steven Anton, Andreas Klintberg, GrowthJeff, Javad Noorbakhsh, Matthew Begun, Ravin Kumar, Sean Hussey, and Skipper Seabold. *CamDavidsonPilon/lifelines: 0.26.0*. Zenodo, 2021.

Alexis Diamond and Jasjeet S. Sekhon. Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies. *Review of Economics and Statistics*, 95(3):932–945, 2013.

D. Faraggi and R. Simon. A neural network model for survival data. *Statistics in medicine*, 14(1):73–82, 1995.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between Optimal Transport and MMD using Sinkhorn Divergences, 2018. URL http://arxiv.org/pdf/1810.08278v1.

J. A. Foekens, H. A. Peters, M. P. Look, H. Portengen, M. Schmitt, M. D. Kramer, N. Brünner, F. Jänicke, M. E. Meijer-van Gelder, S. C. Henzen-Logmans, W. L. van Putten, and J. G. Klijn. The urokinase system of plasminogen activation and prognosis in 2780 breast cancer patients. *Cancer research*, 60(3):636–643, 2000. URL https://pubmed.ncbi.nlm.nih.gov/10676647/.

Thomas R. Frieden. Evidence for Health Decision Making - Beyond Randomized, Controlled Trials. *The New England journal of medicine*, 377(5):465–475, 2017.

Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning Generative Models with Sinkhorn Divergences, 2017. URL https://arxiv.org/pdf/1706.00292.pdf.

Michael F. Gensheimer and Balasubramanian Narasimhan. A scalable discrete-time survival model for neural networks. *PeerJ*, 7:e6257, 2019.

Arthur Gretton, Karsten M Borgwardt, and Malte J Rasch. A Kernel Two-Sample Test. *Journal ofMachine Learning Research 13*, pages 723–773, 2012.

Frank E. Harrell. Evaluating the Yield of Medical Tests. *JAMA: The Journal of the American Medical Association*, 247(18):2543, 1982.

Jennifer L. Hill. Bayesian Nonparametric Modeling for Causal Inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

Paul W. Holland. Statistics and Causal Inference. *Journal of the American Statistical Association*, 81(396):945, 1986.

Chen Huang, Chen Change Loy, and Xiaoou Tang. Local Similarity-Aware Deep Feature Embedding. *Advances in Neural Information Processing Systems*, 29, 2016.

Kosuke Imai and Marc Ratkovic. Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):243–263, 2014. URL https://rss.onlinelibrary.wiley.com/doi/10.1111/rssb.12027.

Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift, 2015. URL https://arxiv.org/pdf/1502.03167.pdf.

Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. Random survival forests. *The Annals of Applied Statistics*, 2(3), 2008.

Ron D. Jachimowicz, Wolfram Klapper, Gunther Glehr, Horst Müller, Heinz Haverkamp, Christoph Thorns, Martin Leo Hansmann, Peter Möller, Harald Stein, Thorsten Rehberg, Bastian von Tresckow, H. C. Reinhardt, Peter Borchmann, Fong Chun Chan, Rainer Spang, David W. Scott, Andreas Engert, Christian Steidl, Michael Altenbuchinger, and Andreas Rosenwald. Gene expression-based outcome prediction in advanced stage classical Hodgkin lymphoma treated with BEACOPP. *Leukemia*, 35 (12):3589–3593, 2021.

Fredrik D. Johansson, Uri Shalit, and David Sontag. Learning Representations for Counterfactual Inference, 2016. URL http://arxiv.org/pdf/1605.03661v3.

Fredrik D. Johansson, Uri Shalit, Nathan Kallus, and David Sontag. Generalization Bounds and Representation Learning for Estimation of Potential Outcomes and Causal Effects, 2020. URL https://arxiv.org/pdf/2001.07426.pdf.

L. Kantorovitch. On the Translocation of Masses. *Management Science*, 5(1):1–4, 1958. URL http://www.jstor.org/stable/2626967.

Jared Katzman, Uri Shaham, Jonathan Bates, Alexander Cloninger, Tingting Jiang, and Yuval Kluger. DeepSurv: Personalized Treatment Recommender System Using A Cox Proportional Hazards Deep Neural Network. *BMC Medical Research Methodology*, 18(1):1, 2018. URL http://arxiv.org/pdf/1606.00931v3.

Gary King and Richard Nielsen. Why Propensity Scores Should Not Be Used for Matching. *Political Analysis*, 27(4):435–454, 2019.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization, 2014. URL http://arxiv.org/pdf/1412.6980v9.

Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal M. Henne. Controlled experiments on the web: survey and practical guide. *Data Mining and Knowledge Discovery*, 18(1):140–181, 2009.

Anders Krogh and John Hertz. A Simple Weight Decay Can Improve Generalization. In J. Moody, S. Hanson, and R. P. Lippmann, editors, *Advances in Neural Information Processing Systems*, volume 4. Morgan-Kaufmann, 1992.

Håvard Kvamme and Ørnulf Borgan. Continuous and Discrete-Time Survival Prediction with Neural Networks, 2019. URL https://arxiv.org/pdf/1910.06724.

Changhee Lee, William Zame, Jinsung Yoon, and Mihaela van der Schaar. DeepHit: A Deep Learning Approach to Survival Analysis with Competing Risks. 2018.

Sheng Li and Yun Fu. Matching on Balanced Nonlinear Representations for Treatment Effects Estimation. *Advances in Neural Information Processing Systems*, 30, 2017.

Richard Liaw, Eric Liang, Robert Nishihara, Philipp Moritz, Joseph E. Gonzalez, and Ion Stoica. Tune: A Research Platform for Distributed Model Selection and Training, 2018. URL https://arxiv.org/pdf/1807.05118.

Danni Lu, Chenyang Tao, Junya Chen, Fan Li, Feng Guo, and Lawrence Carin. Reconsidering Generative Objectives For Counterfactual Reasoning. *Advances in Neural Information Processing Systems*, 33:21539–21553, 2020.

Scott Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions, 2017. URL https://arxiv.org/pdf/1705.07874.

Torben Martinussen and Thomas H. Scheike. *Dynamic Regression Models for Survival Data*. Springer New York, New York, NY, 2006. ISBN 978-0-387-20274-7.

Alfred Müller. Integral Probability Metrics and their enerating classes of functions. 1991.

Vinod Nair and Geoffrey E Hinton. Rectified Linear Units Improve Restricted Boltzmann Machines. 2010. URL https://www.cs.toronto.edu/~fritz/absps/reluICML.pdf.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library, 2019. URL https://arxiv.org/pdf/1912.01703.

Judea Pearl. *Causality: Models, reasoning and interference*. Cambridge University Press, New York, 8. pr edition, 2008. ISBN 9780521773621.

Judea Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 3(0):96–146, 2009.

Judea Pearl. Detecting Latent Heterogeneity. *Sociological Methods & Research*, 46(3):370–389, 2017.

Sebastian Pölsterl. scikit-survival: A Library for Time-to-Event Analysis Built on Top of scikit-learn. *Journal of Machine Learning Research*, 21(212):1–6, 2020. URL https://jmlr.org/papers/v21/20-729.html.

Aaditya Ramdas, Nicolás Trillos, and Marco Cuturi. On Wasserstein Two-Sample Testing and Related Families of Nonparametric Tests. *Entropy*, 19(2):47, 2017.

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

Donald B Rubin. Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, (Vol. 66, No. 5):688–701, 1974.

Claudia Schmoor, Manfred Olschewski, and Martin Schumacher. Randomized and non-randomized patients in clinical trials: Experiences with comprehensive cohort studies. *Statistics in Medicine*, 15(3):263–271, 1996. URL https://onlinelibrary.wiley.com/doi/10.1002/(SICI)1097-0258(19960215)15:3%3C263::AID-SIM165%3E3.0.CO;2-K.

M. Schumacher, G. Bastert, H. Bojar, K. Hübner, M. Olschewski, W. Sauerbrei, C. Schmoor, C. Beyerle, R. L. Neumann, and H. F. Rauschecker. Randomized 2 x 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. German Breast Cancer Study Group. *Journal of Clinical Oncology*, 12(10):2086–2093, 1994.

Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms, 2016. URL http://arxiv.org/pdf/1606.03976v5.

Bharath K. Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert R. G. Lanckriet. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6(none), 2012.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research 15*, pages 1929–1958, 2014. URL https://jmlr.org/papers/volume15/srivastava14a/srivastava14a.pdf.

Annette M. Staiger, Michael Altenbuchinger, Marita Ziepert, Christian Kohler, Heike Horn, Michael Huttner, Katrin S. Hüttl, Gunther Glehr, Wolfram Klapper, Monika Szczepanowski, Julia Richter, Harald Stein, Alfred C. Feller, Peter Möller, Martin-Leo Hansmann, Viola Poeschel, Gerhard Held, Markus Loeffler, Norbert Schmitz, Lorenz Trümper, Tobias Pukrop, Andreas Rosenwald, German Ott, and Rainer Spang. 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(2):543–552, 2020.

R. Tibshirani. The lasso method for variable selection in the Cox model. *Statistics in medicine*, 16(4):385–395, 1997.

Stefan Wager and Susan Athey. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests, 2017.

Xiaonan Xue, Xianhong Xie, Marc Gunter, Thomas E. Rohan, Sylvia Wassertheil-Smoller, Gloria Y. F. Ho, Dominic Cirillo, Herbert Yu, and Howard D. Strickler. Testing the proportional hazards assumption in case-cohort analysis. *BMC Medical Research Methodology*, 13(1):88, 2013. URL https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/1471-2288-13-88.

Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. Representation Learning for Treatment Effect Estimation from Observational Data. *Advances in Neural Information Processing Systems*, 31, 2018.

Jinsung Yoon, James Jordon, and Mihaela van der Schaar. GAN-ITE: Estimation of Individualized Treatment Effects using Generative Adversarial Nets. *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=ByKWUeWA-.

Helena U. Zacharias, Michael Altenbuchinger, Ulla T. Schultheiss, Johannes Raffler, Fruzsina Kotsis, Sahar Ghasemi, Ibrahim Ali, Barbara Kollerits, Marie Metzger, Inga Steinbrenner, Peggy Sekula, Ziad A. Massy, Christian Combe, Philip A. Kalra, Florian Kronenberg, Bénédicte Stengel, Kai-Uwe Eckardt, Anna Köttgen, Matthias Schmid, Wolfram Gronwald, and Peter J. Oefner. A Predictive Model for Progression of CKD to Kidney Failure Based on Routine Laboratory Tests. *American journal of kidney diseases : the official journal of the National Kidney Foundation*, 2021. URL https://pubmed.ncbi.nlm.nih.gov/34298143/.

Supplement

S1. Simulations

To find the best set of model parameters for the three simulations, we employed a comprehensive hyper-parameter grid-search over the listed in Table S1. For each combination, we fitted 50 initializations with randomized 60/40-train/validation splits, which resulted in the presented 600 to 2400 training samples. To avoid over-fitting we used early-stopping based on non-improved validation loss over 50 consecutive epochs for all of the deep neural network recommendation systems ((T-)DeepSurv, SurvITE, (B)ITES). The best set of hyper-parameters is determined based on the minimal mean average C-index evaluated on the validation set. All presented results are based on an independent set containing 1000 samples, respectively, for each simulation. To reduce the computational cost of finding optimal IPM parameters (α and ϵ), we used the best set of hyper-parameters obtained by the corresponding model without IPM regularization ($\alpha = 0$).

Table S1. List of parameters used for the hyper-parameter grid search for the three simulation studies.

HYPER-PARAMETERS	COX	RSF	T-DEEPSURV DEEPSURV	SURVITE	ITES BITES
LAYERS/SHARED LAYERS	-	-	$\{[15, 10, 5], [10, 5]\}$	$\{[50, 50], [20, 20]\}$	$\{[15], [15, 10]\}$
INDIVIDUAL LAYERS	-	-	-	$\{[50, 50], [10, 10]\}$	$\{[10, 5], [5]\}$
LEARNING RATE	$\{0.1\}$	-	$\{0.001\}$	$\{0.001\}$	$\{0.001\}$
BATCH SIZE	-	-	$\{all\}$	$\{300, all\}$	$\{all\}$
l_2-REGULARIZATION	$\{0.1, 0.3, 0.5, 0.7\}$	-	$\{0.01, 0.1, 1\}$	$\{0.1, 0.01, 0.001\}$	$\{0.1, 0.01, 0.001\}$
l_1-REGULARIZATION	$\{0.01, 0.1, 1\}$	-	-	-	-
DROPOUT-RATE	-	-	$\{0.1, 0.3\}$	$\{0.1\}$	$\{0.1, 0.3\}$
IPM STRENGTH α	-	-	-	$\{0, 0.01, 0.1, 1\}$	$\{0, 0.01, 0.1, 1\}$
SINKHORN INTERPOLATION ϵ	-	-	-	-	$\{0.05, 0.1\}$
NUMBER OF TREES	-	$\{1000\}$	-	-	-
MIN SAMPLES SPLIT/LEAF	-	$\{[6, 3], [12, 6]$ $[24, 12]\}$	-	-	-

For SurvITE, we followed the provided example (`https://github.com/chl8856/survITE`). Accordingly, we considered an over-parametrized architecture and scaled the outcome times to obtain 30 discrete time points. For IPM regularization ($\alpha \neq 0$), we used the predefined Wasserstein-distance. Figure S1 shows the achieved Precision in Estimation of Heterogenous Effect (PEHE) for each of the simulations. The obtained results show similar behaviour as the fraction of correctly assigned treatment decisions presented in the main article (Figure 2(d) to 2(f)).
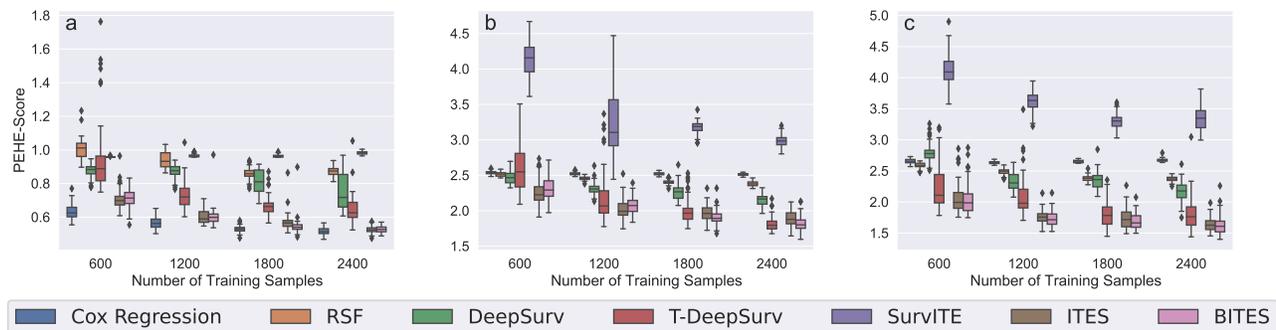

Figure S1. PEHE score obtained for the (a) linear, (b) non-linear and (c) non-linear treatment biased simulation. The boxplots give the distribution of PEHE-scores for 50 consecutive model initializations on independent test data using the best set of hyper-parameters.

S2. Application: RGBSG

For the presented breast cancer application, based on data generated by the Rotterdam and the German Breast Cancer Study Group (RGBSG), we performed a hyper-parameter grid-search for the Cox, RSF and SurvITE treatment recommendation systems with parameters shown in S2. For the remaining models, including (T-)DeepSurv and (B)ITES, we used the *ray[tune]* python package (`https://docs.ray.io/en/latest/tune/index.html`) for the hyper-parameter search. This speeds up computation significantly by

Table S2. List of parameters used for the hyper-parameter search on the RGBSG training data.

| HYPER-PARAMETERS	COX	RSF	T-DEEPSURV	SURVITE	(B)ITES
LAYERS/SHARED LAYERS	-	-	$\{[7,5]\}$	$\{[50,50]\}$	$\{[7,5]\}$
INDIVIDUAL LAYERS	-	-	-	$\{[50,50],[10,10]\}$	$\{[[5,3],[3]]\}$
LEARNING RATE	$\{0.1\}$	-	$[0.0001,0.1]$	$\{0.001\}$	$[0.0001,0.1]$
BATCH SIZE	-	-	$\{all\}$	$\{300\}$	$\{all\}$
l_2-REGULARIZATION	$\{0.3,0.5,0.7,0.9\}$	-	$[0.01,0.1]$	$\{0.0,0.01,0.1,0.5\}$	$[0.01,0.1]$
l_1-REGULARIZATION	$\{0.1,0.5,1\}$	-	-	-	-
DROPOUT-RATE	-	-	$[0.1,0.2]$	$\{0.1\}$	$[0.1,0.2]$
IPM STRENGTH α	-	-	-	$\{0.001,0.1,1,10\}$	$\{0.001,0.01,0.1,1,10\}$
SINKHORN INTERPOLATION ϵ	-	-	-	-	$\{0.05,0.1\}$
NUMBER OF TREES	-	$\{100\}$	-	-	-
MIN SAMPLES SPLIT/LEAF	-	$\{[6,3],[12,6]$ $[24,12]\}$	-	-	-

using a scheduled hyper-parameter optimization. In this case, we used a grid-search for structural parameters and the Sinkhorn parameters, and allowed for random choices for learning rate, l_2 regularization and Dropout rate (paameters given in square brackets on Table S2. For all models, we used 10 reinitializations with randomly drawn 80/20-train/validation splits. This yields 1236 training and 309 validation samples. Similar to the simulation studies, we avoided over-fitting by using early-stopping if the validation loss did not improve within 50 consecutive epochs epochs. The final models were selected by the minimal validation loss achieved for all of the hyper-parameter combinations and reinitializations. This model was then evaluated for an independent test cohort of 686 patients given by the GBSG Trial 2, with results shown in Figure 3, Figure S2 and Table 1.
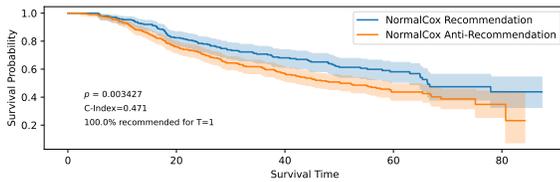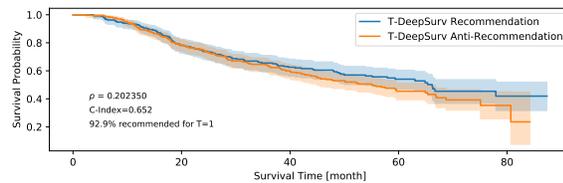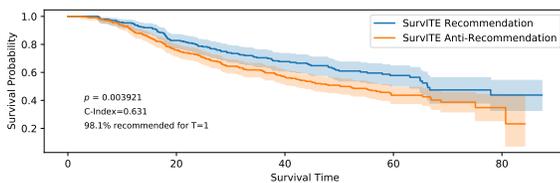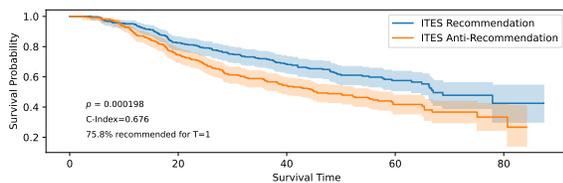

(a) RGBSG for Cox regression.
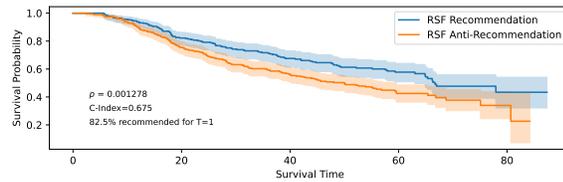
(b) RGBSG for RSF.
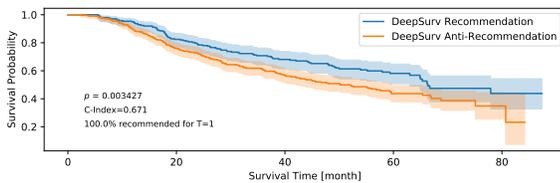
(c) RGBSG for DeepSurv.

(d) RGBSG for T-DeepSurv.

(e) RGBSG for SurvITE.

(f) RGBSG for ITES.

Figure S2. Kaplan Meier curves corresponding to Figure 3, for (a) Cox regression, (b) RSF, (c) DeepSurv, (d) T-DeepSurv, (e) SurvITE and (f) ITES. Each of the plots contains the p-value comparing the recommended and anti-recommended group, the obtained C-index and the fraction of patients that the algorithm recommends to administer the treatment.