A Hierarchical Bayesian Model for Personalized Survival Predictions

Alexis Bellot, Mihaela van der Schaar, *Fellow, IEEE*

***Abstract*—We study the problem of personalizing survival estimates of patients in heterogeneous populations for clinical decision support. The desiderata are to improve predictions by making them personalized to the patient-at-hand, to better understand diseases and their risk factors, and to provide interpretable model outputs to clinicians. To enable accurate survival prognosis in heterogeneous populations we propose a novel probabilistic survival model which flexibly captures individual traits through a hierarchical latent variable formulation. Survival paths are estimated by jointly sampling the location and shape of the individual survival distribution resulting in patient-specific curves with quantifiable uncertainty estimates. An understanding of model predictions is paramount in medical practice where decisions have major social consequences. We develop a *personalized interpreter* that can be used to test the effect of covariates on each individual patient, in contrast to traditional methods that focus on population average effects. We extensively validated the proposed approach in various clinical settings, with a special focus on cardiovascular disease.**

***Index Terms*—Personalized risk prediction, Personalized prognosis, Survival analysis, Bayesian inference, Permutation tests.**

I. Introduction

We focus in this work on developing a class of new prognostic scoring models for time-to-event analysis, such as the time to disease onset or time to death; the developed prognostic models are personalized and interpretable. We are driven by the challenge of capturing the diverse information modern medical data sets provide to inform clinical decision making at a greater level of depth. Evidence from medical research and practice have shown that health care, and particularly data-driven methodologies based on large heterogeneous cohorts, must be specifically tailored to the patient and the disease in question in order to improve prognosis and medical understanding [1], i.e. it must be personalized. Therapies and procedures effective on a particular patient subgroup might result in a very different outcome for a different subgroup [2]; thus predictive models not accounting for heterogeneity may perform well on average, but will likely underestimate or overestimate the risks for specific sub-populations. For this reason the field of personalized medicine has emerged as a promising approach to understanding survival and the onset of diseases [1] at the individual level with the potential to significantly improve health care [3].

The prognostic scoring method developed in this work is general and can be applied to a variety of diseases and medical conditions. However to illustrate the need for a flexible and personalized approach to prognostic scoring, we use a cohort of critical patients wait-listed to receive a heart transplant. Patients on the wait-list are subject to an extremely volatile status; available treatments, co-morbidities and past history are known to have a very heterogeneous effect on individual survival outcomes, which makes predicting and minimizing mortality in the wait-list a notoriously difficult problem, but crucial for an effective treatment allocation scheme [4]–[6].

Predictive models describing survival dynamics fall under the umbrella of survival analysis. Broadly defined, survival analysis is a method for analyzing data where the target variable is the time to the occurrence of an event of interest. The goal of these models is not only to examine the effects on the time until an event occurs, but also to assess the relationship of survival time to explanatory variables. Survival and event history analysis departs fundamentally from common supervised learning problems by not only focusing on the outcome but also analysing the time to an event. Widely used regression methods for survival analysis include Cox-proportional hazards model [7] and parametric models based on Weibull, Exponential and Log-normal distributions [8], these provide interpretable variable effects but inference in individual patients is restricted by a common survival shape. More flexible non-parametric methods have also been proposed, for instance, tree based methods such as Random Survival Forests (RSF) [9], Conditional Inference Forests [10] and Bayesian Additive Regression Trees [11]–[13] model survival times solely data-driven, without making any distributional assumptions. These have been shown in [14] to achieve performance gains in heterogeneous medical cohorts. However, differences in model outputs for two different patients or the impact of covariates cannot be inferred from these models, and thus clinicians are unable to explain decisions based on them, which is critical in order to be clinically useful [15], [16].

Contribution. In this paper, we regard individual survival trajectories as being generated by a hierarchical process in which heterogeneity is captured in all aspects of the survival distribution, and subsequently propagated through the hierarchy resulting in *personalized* survival distributions. The proposed model is Bayesian: we assign a prior distribution over the space of latent parameters of a parametric survival distribution, and update the posterior distribution given

A.Bellot and M. van der Schaar are with the University of Oxford. E-mail: alexis.bellot@eng.ox.ac.uk

Manuscript submitted January 15, 2018

observations from medical data. Uncertainty in predictions is directly quantifiable from the posterior distribution over individual survival curves, whose parameters relate to actual clinical variables. Model outputs can thus be interpreted in the context of the clinical application at hand. The influence of covariates on survival is encoded in latent variables through a non-parametric probabilistic model; high-dimensional feature spaces and weak signals are then efficiently represented by letting the data determine the model complexity. We provide a detailed description of the model in Section III, and then propose a sampling algorithm for learning posterior distributions of parameters in Section IV.

We supplement our clinical decision support tool with a post-processing procedure we call the *personalized interpreter* to formally test for the significance of covariates on survival outcomes for an individual patient. Our aim is to infer what covariates *significantly* influence survival for the patient-at-hand, in a personalized manner. Henceforth, we refer to our proposed clinical prognostic scoring model as Hierarchical Bayesian Survival (HBS) model.

II. Related Work

Predictive models. Closest to our work are techniques which model survival with a regression function embedded in a parametric model, this was considered in [17] and [18]. The work in [17] considers a hierarchical Bayesian Weibull linear model to identify relevant biomarkers through shrinkage priors. While they are able to perform variable selection, for prediction in heterogeneous cohorts, the Weibull and linearity of covariate effects assumptions are restrictive. In [18], the authors use Bayesian Additive Regression Trees within a Weibull, Log-Normal and Cox-Proportional Hazards model, but limit the covariate influence on the location of the distribution which fails to model sub-population dynamics in heterogeneous cohorts. In this paper, we depart from these works by explicitly avoiding a common survival shape for the whole population and specify a more general survival distribution. We are therefore able to learn a richer set of hazard shapes from data. This is useful for personalized medicine in that the random shape parameters can be viewed as characteristic to a particular subgroup, that may vary across subgroups due to biological or environmental differences that are measurable [19]. In [13], the authors apply Bayesian Additive Regression Trees (BART) [11] directly to survival data by discretizing the time horizon and modelling survival status in narrow time windows.

Understanding variable effects – Interpreter. Techniques inferring variable effects aim at understanding predictive contributions of individual variables in an attempt to explain model predictions. We focus here on model-agnostic approaches, these are typically related to changes in model outputs after perturbation of inputs or model architecture (see [20] and [21]) or related to fitting a simpler model to the predictions themselves [22]. These works are concerned with population average variable effects, and thus will miss the differences in variable impact within patient populations, giving uninformative explanations in heterogeneous cohorts. Techniques inferring individual variable effect have been proposed in [23] as an extension to partial dependence plots [24] by evaluating model predictions fixing the set of covariates relating to a particular individual rather that averaging out their impact over the covariate distribution. Uncertainty in the model construction, however, is not captured by these methods - a threshold for significance is not discussed. We propose a permutation-based statistical test that reconstructs a null distribution capturing the uncertainty in individual covariate effects and thus we are able to assess the statistical significance of covariates on an individual basis.

III. Methods

A. Problem formulation

We are concerned with heterogeneous patients populations whose covariates may influence their survival outcome in potentially very different ways. Each patient i is characterized by a d-dimensional vector of covariates $X_i \in \mathcal{X}$ (with realization $\boldsymbol{x}_i$), an outcome variable $T_i \in \mathbb{R}^+$, the time until the event of interest drawn from a distribution $T_i \sim \mathbb{P}(.|X_i)$ and an indicator variable δ_i referring to the type of event observed. Patients being followed in a study may drop-out resulting in a potential event being unobserved; thus in this case δ_i refers to right censoring ($\delta_i = 0$) or the occurrence of the event ($\delta_i = 1$).

Our goal is to estimate the survival function $S : (\mathcal{X}, \mathcal{T}) \rightarrow [0,1]$ which represent the probability of event occurrence after time t as a function of time t and patient covariates X_i,

$$S(t|X_i) = \mathbb{P}(T_i > t|X_i) \tag{1}$$

As discussed in the Introduction, the relationship between patient covariates, time and survival outcome will be complex for many modern data sets. Specifically we aim to,

- Allow for flexible interactions between patient covariates, time and survival that are personalized to each individual. That is, $\mathcal{S}$ will be modelled by a flexible function with few assumptions constraining its behaviour.
- Provide a tractable prognostic model whose features remain clinically meaningful (relate to actual clinical variables) and which provides understanding of disease progression within patient sub-populations.

The relationship between survival and patient covariates is to be estimated from an observational data set $\mathcal{D}$ comprising n independent samples of the random tuple $\{X_i, \delta_i, \delta_i T_i + (1-\delta_i)C_i\}$. C_i represents the censoring time drawn from a distribution assumed independent from covariates X_i, a common assumption in the survival literature.

The probability of event within a suitable time window Δt, $\mathbb{P}(T > t + \Delta t | T > t, X_i)$ is of great importance; for instance, it is used as a risk score based on which clinicians prioritize therapies for patients and design treatment plans. We define

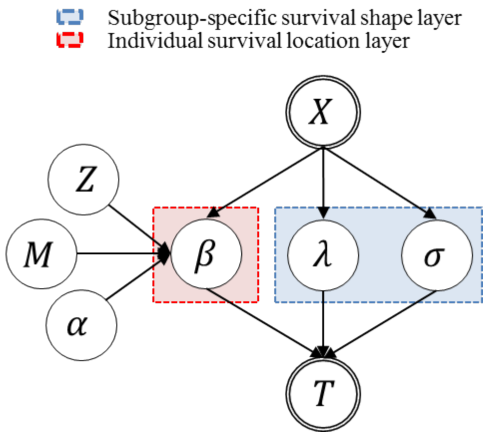
Subgroup-specific survival shape layer
Individual survival location layer
X
Z
M
α
β
λ
σ
T

Fig. 1. Graphical Model of the proposed HBS. Double-circled variables are observed.

the probability density function $f_i(t) = -\partial S_i(t)/\partial t$. We can equivalently characterize survival through the hazard function defined as

$$\lambda(t|X_i) = \lim_{dt \to 0} \frac{\mathbb{P}(t \leq T_i \leq t + dt | T_i \geq t, X_i)}{dt} \tag{2}$$

which represents the instantaneous risk of the event occurring, and thus gives insight into the rate of change of survival probabilities. Survival and hazard functions are related by $\lambda_i(t) = -\partial \log(S_i(t))/\partial t$.

B. Hierarchical Bayesian Model

In this section we introduce a patient-specific survival prediction model with the goal of making only few assumptions about structural aspects of the data while remaining interpretable by fully specifying the survival and hazard functions. Through a generative probabilistic model, we estimate the survival distribution T_i directly and model its components with separate models to encode different aspects of a heterogeneous population.

1) Time-to-event model: The time to the event of interest, such as diagnosis of a disease or death, T_i, $i = 1...n$ is assumed to be generated from a Generalized Gamma distribution ($\mathcal{GG}$) [25].

$$T_i \sim \mathcal{GG}(\beta_i, \sigma_i, \lambda_i) \tag{3}$$

The motivation is that it contains as special cases most of the familiar distributions used in time-to-event settings such as the Weibull ($\lambda = 1$), Gamma ($\sigma = \lambda$), Log-Normal ($\lambda = 0$) and Exponential ($\sigma = \lambda = 1$) distributions and is thus able to flexibly model a wide range of processes while its parameters relate to meaningful real-world quantities such as the hazard shape (which is unavailable in purely nonparametric models). Formally, a random variable $T \in \mathbb{R}^+$ is $\mathcal{GG}(\beta, \sigma, \lambda)$ distributed if its probability density function, for $t > 0$, is of the following form:

$$f(t) = \frac{|\lambda|(\lambda^{-2})^{\lambda^{-2}}}{\sigma t \Gamma(\lambda^{-2})} (e^{-\beta} t)^{1/\sigma\lambda} \exp\{-\lambda^{-2}(e^{-\beta} t)^{\lambda/\sigma}\}$$

where $\Gamma(x)$ denotes the Gamma function. The parameters $(\beta, \sigma, \lambda) \in \mathbb{R} \times \mathbb{R}^+ \times \mathbb{R}$ model the location, scale and shape of the distribution respectively. β acts multiplicatively on time only; thus for fixed parameters (σ, λ), β governs the median time-to-event time, i.e. $\beta = \log(median) + c(\sigma, \lambda)$, ($c$ a function independent of β). This makes parameter β a natural candidate to express the influence of covariates.

We introduce prior distributions on all latent parameters to construct the hierarchical process from which the observed data is assumed to be generated. The graphical model in Figure 1 illustrates this structure. Motivated by ensemble methods we leverage the relationship between the functional terms (β, σ, λ), each optimizing a specific aspect of the time-to-event distribution to model heterogeneity. For instance, the distributional shape is subgroup-specific while the median time-to-event is unique to each individual.

2) Regression model: We treat β_i as a latent variable that will incorporate the influence of a patient's covariates on its expected median survival. To accommodate for the wide variability in individual features and their impact on observed survival dynamics, β_i is represented with Bayesian Additive Regression Trees. BART is a probabilistic "sum of trees" model with normal noise in which individual trees h are prevented from overwhelming the fit by imposing a regularization prior on the tree structure $\boldsymbol{Z} = \{Z_j\}_j$ and leaf output $\boldsymbol{M} = \{M_j\}_j$ parameters. We refer the reader to [11] for a more detailed exposition. Formally we write,

$$\beta_i | \boldsymbol{x}_i \sim g(\boldsymbol{x}_i) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \alpha) \tag{4}$$

$$g(\boldsymbol{x}_i) := \mathbb{E}\left(\sum_j h(\boldsymbol{x}_i; Z_j, M_j)\right) \tag{5}$$

$$\alpha \sim 1/\chi^2 \tag{6}$$

We set the prior for α to be the conjugate inverse chi-square distribution [11]. g represents the BART hypothesis for the relationship between covariates and median survival which specifies the prior mean of the $\boldsymbol{\beta}$ latent variables thereby encouraging sampling close to the estimates of g. The error in the estimation determines the confidence in the estimated relationship and thus regulates how certain the updated prior is. This prior distribution of $\boldsymbol{\beta}$ will itself be updated within the sampling algorithm to jointly optimize both the prior and posterior given observed covariates according to highest likelihood of the observed events and time to event.

This structure allows us to naturally incorporate complex regression functions to model the influence of covariates and *inform* the prior with this relationship (prior mean) and confidence in predictions of the complex regression function (prior variance).

The choice of tree-based methods is especially appealing in contexts of heterogeneous populations. Tree based methods have the advantage of providing predictions based on multiple features without prior selection, we are able to recover sparse signals in high dimensional feature spaces (e.g. in genetics). They are robust, that is, not affected by monotonic transformations such as scaling or shifting of the data. This can be important as many covariates are recorded subject to nuisance variability which may bias results [26].

Patients in heterogeneous populations often display fundamentally different survival trajectories beyond differences in median event times that can be discovered from data. As an example, severe diseases such as breast cancer often require invasive treatments that may significantly increase the short term risk of death but lower it in the long run relative to a control population, so that hazard shapes are different. Methodologies not accounting for this heterogeneity –which although may work well on average– will likely provide inaccurate estimates for large parts of the population (”one size does not fit all”). To deal with this heterogeneity, we allow for different behaviours to be learned effectively from data through the dependence of subgroup-specific shape parameters on a partition of interest $\{\mathcal{C}_j\}_{j=1}^J$. We specify

$$\sigma_i = \sigma_{(j)} I(i \in \mathcal{C}_j), \quad \sigma_{(j)} \sim \mathcal{G}(\eta_0, \eta_1) \tag{7}$$

$$\lambda_i = \lambda_{(j)} I(i \in \mathcal{C}_j), \quad \lambda_{(j)} \sim \mathcal{N}(\gamma_0, \gamma_1) \tag{8}$$

Individuals within the same subgroup share the survival shape specification. $\mathcal{G}$ denotes the Gamma distribution and $(\eta_0, \eta_1, \gamma_0, \gamma_1)$ are fixed hyper-parameters. Prior distributions are chosen by sampling to be weakly informative about the survival shapes; we generate parameter values (β, σ, λ) from the priors and ensure the corresponding survival curves $\mathcal{GG}(\beta, \sigma, \lambda)$ cover a broad set of possible survival dynamics. The partition $\{\mathcal{C}_j\}_{j=1}^J$, in practice, will be provided by field experts according to the sub-populations to be analyzed or depending on context-based meaningful criteria. The influence of co-morbidities on survival, in particular, can be analyzed in our framework by specifying different shape parameters to the relevant patient sub-populations, and thus quantitatively and qualitatively compare the sub-populations with different co-morbidities. An example of the use of our model for multi-morbid populations is given in section V-C.

IV. Learning and Inference

We infer optimal latent parameter variables $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\sigma}, \boldsymbol{\lambda}, \alpha, \boldsymbol{Z}, \boldsymbol{M})$ from the posterior distribution of parameters given the data and model. The posterior summarizes the distributional hypothesis about the data generating process that most closely agrees with the observed data and prior model. We compute the posterior with Bayes formula.

$$p(\boldsymbol{\theta}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{\theta}) p(\boldsymbol{\theta})$$

and thus under the assumption that model parameters are related by the dependency structure in Figure 1 the joint posterior is given by,

$$p(\boldsymbol{\beta}, \boldsymbol{\sigma}, \boldsymbol{\lambda}, \alpha, \boldsymbol{Z}, \boldsymbol{M}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{\beta}, \boldsymbol{\sigma}, \boldsymbol{\lambda}) p(\alpha) p(\boldsymbol{Z}, \boldsymbol{M}) \times \prod_j p(\sigma_{(j)}) p(\lambda_{(j)}) \prod_{i \in \mathcal{C}_j} p(\beta_i | \boldsymbol{Z}, \boldsymbol{M}, \alpha) \tag{9}$$

Algorithm 1: Posterior Distribution Sampling

```
Input: Dataset $\mathcal{D}$, number of iterations $T$, prior
       distributions for $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\sigma}, \boldsymbol{\lambda}, \alpha, \boldsymbol{Z}, \boldsymbol{M})$.
Initialize $\boldsymbol{\theta}^{(0)} = (\boldsymbol{\beta}^{(0)}, \boldsymbol{\sigma}^{(0)}, \boldsymbol{\lambda}^{(0)}, \alpha^{(0)}, \boldsymbol{Z}^{(0)}, \boldsymbol{M}^{(0)})$;
for t from 1 to T do
    • $(\boldsymbol{Z}^{(t)}, \boldsymbol{M}^{(t)}) \leftarrow$ update $(\boldsymbol{Z}, \boldsymbol{M})$ such that
      $g : (\mathcal{X}, \boldsymbol{Z}, \boldsymbol{M}) \rightarrow \boldsymbol{\beta}^{(t-1)}$ is optimal, estimated
      with BB-MCMC;
    • $\alpha^{(t)} \leftarrow$ sample from full conditional
      $p(\alpha|\boldsymbol{Z}^{(t)}, \boldsymbol{M}^{(t)}, \boldsymbol{\beta}^{(t-1)})$;
    • Update prior means $\mathbb{E}\left(\boldsymbol{\beta}^{(t-1)}\right) := g(\boldsymbol{X}; \boldsymbol{Z}^{(t)}, \boldsymbol{M}^{(t)})$
      and variance Var$\left(\boldsymbol{\beta}^{(t-1)}\right) := \alpha^{(t)}$;
    • for i from 1 to N do
        $\beta_i^{(t)} \leftarrow$ sample from Markov chain with target
        $p(\beta_i|\boldsymbol{\beta}_{-i}^{(t-1)}, \boldsymbol{\sigma}^{(t-1)}, \boldsymbol{\lambda}^{(t-1)}, \boldsymbol{Z}^{(t)}, \boldsymbol{M}^{(t)}, \mathcal{D})$;
      end
    • for j from 1 to J do
        $\sigma_j^{(t)} \leftarrow$ sample from Markov chain with target
        $p(\sigma_j|\boldsymbol{\beta}^{(t)}, \boldsymbol{\sigma}_{-j}^{(t-1)}, \boldsymbol{\lambda}^{(t-1)}, \mathcal{D})$;
      end
    • for j from 1 to J do
        $\lambda_j^{(t)} \leftarrow$ sample from Markov chain with target
        $p(\lambda_j|\boldsymbol{\beta}^{(t)}, \boldsymbol{\lambda}_{-j}^{(t-1)}, \boldsymbol{\sigma}^{(t)}, \mathcal{D})$;
      end
end
Output: Approximate samples $\{\boldsymbol{\theta}^{(t)}\}_{t=1}^T$ from $p(\boldsymbol{\theta}|\mathcal{D})$
```

The likelihood of observed data is given by,

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_j \prod_{i \in \mathcal{C}_j} f(t_i; \beta_i, \sigma_{(j)}, \lambda_{(j)})^{\delta_i} S(t_i; \beta_i, \sigma_{(j)}, \lambda_{(j)})^{1-\delta_i} \tag{10}$$

where $S(t_i) := \mathbb{P}(T > t_i)$ and $\{\mathcal{C}_j\}_{j=1}^J$ denotes the subgroup partition. The expressions involved make direct posterior inference intractable. To approximate the posterior we rely on sampling from a Markov Chain with stationary distribution the posterior in (9) using a Metropolis within Gibbs Markov Chain Monte Carlo (MCMC) scheme [27]. We start by initializing parameter values $(\boldsymbol{\beta}, \boldsymbol{\sigma}, \boldsymbol{\lambda})$ with the maximum likelihood parameter estimates of a Generalized Gamma model for the average patient while the remaining variables are drawn from their prior distributions. In each iteration of the procedure, given the initialized parameter $\boldsymbol{\beta}$, we update the tree structure and terminal node parameters (variables $\boldsymbol{Z}, \boldsymbol{M}$) using a tailored version of the Bayesian backfitting MCMC (BB-MCMC) algorithm [11], [28]. The dispersion parameter α of the error term ϵ_i is then updated by sampling from the full conditional (since we chose a conjugate prior for α). The updated regression tree model g and α parameter determine at this stage the prior mean and variance of $\boldsymbol{\beta}$, $\beta_i \sim \mathcal{N}(g(\boldsymbol{x}_i), \alpha)$. Subsequently, each parameter related to the survival distribution, $(\boldsymbol{\beta}, \boldsymbol{\sigma}, \boldsymbol{\lambda})$, is updated by cycling through the parameter space, each component sequentially. Each marginal conditional distribution is sampled with a Metropolis step, the acceptance ratio is computed from the available fully specified likelihood and prior distribution since we restrict ourselves to independent prior distributions. We log-transform the σ parameter to the real-valued domain and use a random walk proposal distribution for all variables with adaptive step size (proposal variance), updated every 50 iterations to ensure an acceptance rate of around 40%. Algorithm 1 details the complete procedure.

Offline computational complexity for training the HBS prognostic model. We would like to note that at run-time, when the `HBS` model is deployed, the computational complexity is minimal. The learning complexity only plays a role when the model is learned on the basis of the available data. Hence, its complexity does not represent a major issue in practice. However, to understand the complexity associated with training the `HBS` model as compared to other prognostic models, we provide a brief analysis of this offline complexity and provide training times for all experiments. The complexity of computing posterior samples for `HBS` is $\mathcal{O}(T \times (G(N, D) + N + J))$ where N is the number of patients, D is the number of covariates, J is the number of clusters and $G(N, D)$ is the computational complexity of estimating g, which has been shown empirically in [12] to be approximately linear in both N and D. In every iteration of the sampler g is updated. The remaining operations can be performed in $\mathcal{O}(N + J)$.

V. Interpreter: Understanding Personalized Survival Estimates

To understand the variable influence of a covariate on an individual patient we introduce at this stage a *personalized interpreter* to explain the learned relationship from patient covariates to median survival time, $\mathcal{X} \rightarrow \hat{\boldsymbol{\beta}}$. Given the elaborate nature of survival dynamics, this is of clinical interest for updating current consensus guidelines and to generate clinical hypotheses. By looking at individual patients and their similarities we advocate our method as a *phenotype discovery* tool to learn latent representations that capture structure in data and may represent real patterns of illness. The proposed approach is composed of two parts: $1)$ an individual variable effects measure, and $2)$ a statistical test to assess the significance of the predicted individual variable effects.

A. Individual Variable Effects

We approximate the individual covariate impact on survival with Individual Conditional Expectation curves (ICE) [23]. ICE curves are computed from a fitted model by varying the level of the covariate of interest while fixing the remaining patient information. Consider a variable of interest X_s observed to take the value $X_{j,s} = x_{j,s}$ (the (j, s) entry of data matrix $\boldsymbol{X}$) for patient j and a baseline level $X_s = b$ a clinician would want to compare with. For a fitted model g we define the individual impact on survival as:

$$\phi(x_{j,s}; b) = \exp\left(\hat{g}(X_{j,s} = x_{j,s}, \boldsymbol{x}_{j,-s}) - \hat{g}(X_{j,s} = b, \boldsymbol{x}_{j,-s})\right) - 1 \quad (11)$$

where $\boldsymbol{x}_{-s,j}$ denotes all other covariates of patient j which are held fixed. From the relationship between β and median survival in the $\mathcal{GG}$ distribution we can show that for an individual patient,

$$\phi(x_{j,s}; b) = \frac{\widehat{med}(T_j | X_{j,s} = x_{j,s}) - \widehat{med}(T_j | X_{j,s} = b)}{\widehat{med}(T_j | X_{j,s} = b)} \quad (12)$$

This is the relative change in individual predicted median survival time attributed to a value $X_{j,s} = x_{j,s}$. Note that the ICE curve averaged over all patients is Friedman's Partial Dependence curve [24]. A curve can be constructed relative to a baseline b by varying y in $\phi(x_{j,s} = y; b)$.

In this way variable impact can be assessed at the individual level, that is, based on the combination of covariates of an individual patient. This reveals how a treatment, say, applied to the whole population may affect different subgroups and individual patients differently. This is in contrast to traditional methods that infer the impact on an average patient, such as the magnitude of estimated parameters in a Cox model. However, individual estimates might exhibit large variability, dependent on model construction or the specific population being analyzed. Therefore we developed a nonparametric statistical test which evaluates the null hypothesis of no effect on survival of a particular covariate for a particular patient, which frames and quantifies our uncertainty in estimated individual effects.

B. Statistical Significance of Predicted Effect

A high predicted individual covariate effect is not immediately indicative of a corresponding significant covariate effect since variability in the model construction might lead us to conclude that the predicted effect was generated by chance. To assess statistical significance of the patient-specific effect of a covariate on median survival we propose a non-parametric permutation test shown in Algorithm 2. We repeatedly estimate the individual covariate effect (with (12)) for a selected patient after permuting the covariate of interest in the data to obtain a null distribution of *no* effect. The predicted effect is then compared to the quantiles of this null distribution to estimate the significance of the predicted effect. The permutation test accurately represents our process of inference because our null hypothesis is that of no impact of the variable on the outcome. The intuition is that generated individual covariate effects under the hypothesis of *no* effect far from the original predicted effect suggest statistical significance as it would be unlikely to have been generated by chance.

VI. Experiments

In this section we perform a statistical analysis on UNOS observational medical data (see below, section V-B); and evaluate the predictive ability of `HBS` in comparison to the wide range of models described previously. Further predictive performance comparisons on two additional publicly available medical data sets, related to Cardiovascular diseases and Breast Cancer, are provided in the Appendix.

Algorithm 2: Individual Significance Test

```
Input: predicted effect $\phi^{obs}(x_{j,s}; b)$ for patient of interest
       $j$ and number of null samples $n$
for i from 1 to n do
    $\boldsymbol{X}' \leftarrow$ data $\boldsymbol{X}$ with variable $s$ randomly shuffled;
    $\hat{g} \leftarrow$ optimal $g : \boldsymbol{X}' \rightarrow \boldsymbol{\beta}$ ;
    $\phi^{(i)}(x_{j,s}; b) \leftarrow$ individual impact under model $\hat{g}$;
end
$p$-value $\leftarrow |\{i : |\phi^{obs}(x_{j,s}; b)| > |\phi^{(i)}(x_{j,s}; b)|\}|/n$
Output: Significance of predicted effect $\phi^{obs}(x_{j,s}; b)$
```

A. Performance Assessment

Performance assessment of survival data is complicated due to the presence of censoring, labels and event times are not observed for every individual. In this paper we adopt two common approaches used in the literature: the time-dependent concordance index (C-index) [29] defined as,

$$C(t) := \mathbb{P}(\hat{S}_i(t) > \hat{S}_j(t) | \delta_i = 1, t \leq T_j, T_i > T_j) \qquad (13)$$

and the mean square error (MSE), called Brier Score (BS) [30] in this setting:

$$BS(t) := \mathbb{E}[(\delta_i(t) - (1 - \hat{S}_i(t)))^2] \qquad (14)$$

where $\delta_i(t)$ denotes the survival status of patient i at time t (1 if the event has occurred, 0 otherwise).
The time-dependent C-index as defined above corresponds to the probability that predicted survival times are ranked in *accordance* to the actual observed survival times, it thus serves as a measure of the discriminative power of a model. The C-index is defined on the $[0.5, 1]$ interval, with 0.5 corresponding to performance of random guesses and 1 corresponding to perfect ordering of survival times. The Brier Score is a measure of calibration, summarizing the predictive accuracy of a model. It measures the discrepancy between the actual observed status and the probability of event at that time. A low Brier Score is desirable. In all experiments, these metrics are adjusted for censoring, the implementations are done with the functions `cindex` and `pec` from the R package `pec`.

We compare our model with a wide range of baseline algorithms from the Biostatistics and Machine Learning communities.

Biostatistics. Prediction algorithms from the field of Biostatistics are widely used in medical practice due to straightforward implementation and interpretable model outputs. As a first comparison we use the Weibull parametric model (Weibull) [8] with linear predictor under the assumption that the effect of a covariate is to accelerate or decelerate the time to event process by a constant. As a semi-parametric alternative we implement Cox-Proportional Hazard Model (Cox) [7] which assumes instead hazards to be in constant proportion over time but relaxes the distributional assumption of the Weibull model, the baseline hazard function remains unspecified. Its flexibility and ease of interpretation make it widely used for time-to-event analysis in all fields. We evaluate also the fully non-parametric estimator, the Nonparametric Additive Hazards Model of Aalen (Aalen) [31].

Machine Learning. The Machine learning community as recently contributed extensively to developing survival models and have shown performance gains in complex heterogeneous patient cohorts. We consider the tree-based algorithms Random Survival Forest (RSF) [9] and Conditional Inference Forest (CForest) [10] as frequentist fully non-parametric competing methods. We consider also Bayesian Additive Regression Trees (BART) implemented directly [13] as a probabilistic alternative to the frequentist procedures above. We compare with Cox proportional hazards model by component-wise likelihood-based boosting (CoxBoost) [32] that uses an L2-norm penalized version of the partial log-likelihood maximized in Cox with parameter values updated through boosting iterations. Finally we implement the Weibull-Tree model (Wei-Tree) from [18] with their recommended prior specifications.

Hyperparameters are set by cross-validation for Coxboost and BART. Biostatistics techniques do not require hyperparameter specifications and are implemented off the shelve. RSF and CForest are implemented with 500 and 1000 base trees, the default specifications in the implementations of [9] and [10] used here.

B. Data Description

We retrospectively analyzed a cohort from the publicly available United Network for Organ Sharing (UNOS) data set [1]. UNOS encompasses an open cohort of prospectively collected donor specific and follow-up data from 1987 containing all patients undergoing orthotopic heart transplantation in the U.S. The miss-balance of heart transplant demand and available pool donors requires a carefully designed allocation scheme that establishes a priority order to distribute healthy hearts with the goal to maximize the allocation of donor hearts to patients with the greatest likelihood of dying while waiting for a donor organ. While on such a wait-list, Left Ventricular Assist Devices (LVADs), a mechanical pump that is implanted inside a patients' chest to help a weakened heart pump blood, have increasingly been used to bridge patients to orthotopic heart transplantation. Currently, in the circumstances of rapid deterioration, LVAD mechanical failures and infections, patients get assigned high priority for a heart transplant; in normal conditions, due to the frequent complications of these devices, medically stable patients implanted with an LVAD can get high priority status (1A) for a period of 30 days in which transplant occurs [33]. With the introduction of second generation, continuous flow LVADs (cfLVAD), the UNOS criteria for listing patients for heart transplantation and for determining their status for priority has not resulted

[1] Available at `https://www.unos.org/data/`

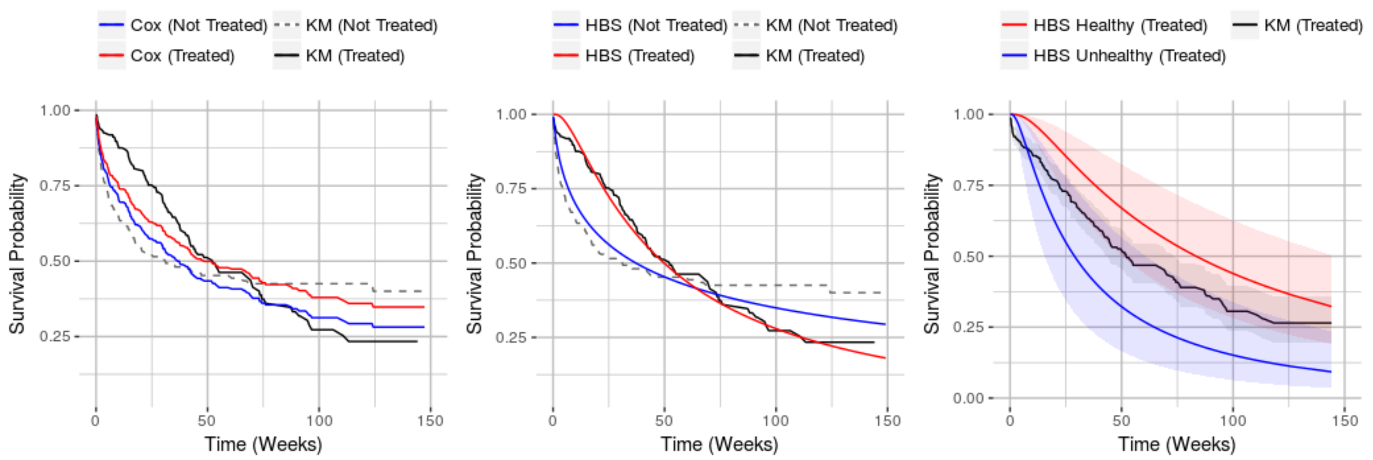

Fig. 2. Survival functions for various models and subgroups in the UNOS data set (best seen in color). **Left Panel,** in red (lighter) and blue (darker) are depicted Cox curves for populations who received a cfLVAD (treated) and those who did not (not treated) respectively together with the empirical KM curves for the two groups. **Middle Panel,** for the same subgroups are shown average survival curves for treated and non-treated populations estimated by `HBS`. **Right Panel,** are shown the KM curve for the treated population together with individual curves predicted for two treated real patients with different characteristics. These estimates are accompanied with 90% credible intervals.

in the best use of a limited resource. The risk of serious complications is now low, which may result in less urgency to perform heart transplantation in clinically stable patients, current patients receiving donor hearts are no longer the ones at greatest risk of mortality [4]–[6]. cfLVADs are associated with improved functional capacity and quality of life on average but its effect is remains highly heterogeneous [34]. The goal of the present study is to determine the effect of cfLVADs on survival and investigate the individual impact of patient characteristics and its interactions with cfLVADs.

For the purposes of this analysis, we selected the cohort entering the wait-list in 2010; this is because patients who received an LVAD in this period were guaranteed to have received a second generation cfLVAD, and have been followed up for at least 6 years to assess their survival trajectories in the wait-list. In addition, paediatric patients, that is, individuals under 21 years old, were left out from the study. The extracted cohort comprises 792 patients described by 17 covariates including demographic characteristics, a number of physiological markers and interventions. Table I shows the feature distribution for the control and cfLVAD treated subpopulations.

TABLE I
THE FEATURE DISTRIBUTION OF THE EXTRACTED COHORT OF UNOS.

Mean (Std. Dev.)	Treated	Control
# of Patients	270	522
Age (y)	52.97 (11.1)	52.24 (12.3)
Weight (kg)	86.16 (20.3)	90.11 (19.1)
Height (cm)	174.45 (10.5)	174.04 (10.0)
Diabetes	37%	34%
Male	70%	72%
Body Mass Index	29.42 (5.3)	28.30 (5.3)
Creatinine (dose)	1.36 (0.7)	1.45 (1.0)
Ventilator	3.3%	3.8%
ECMO	1.48%	1.53%
VAD	20%	4%
Blood Type A	30%	34%
Blood Type B	10%	10%
Blood Type AB	1%	2%
Blood Type 0	59%	54%
IABP	7%	6%
# prev. transplants	0.02 (0.13)	0.07 (0.27)

C. Model specification and Analysis

Parameter values were estimated by `HBS` with 5000 iterations of the MCMC sampler in addition to 1000 iterations as burn-in, which we found to be sufficient for exploration of the posterior distribution space. Posterior convergence was assessed by running 5 chains with random starting points. Prior distributions were chosen to generate prior survival distributions consistent with expected survival dynamics (ie survival times within the range (0,350) weeks and survival shape undefined). We found $\lambda \sim \mathcal{N}(0,2)$, $\sigma \sim \mathcal{G}(1,1)$ and $\beta \sim \mathcal{N}(5,1)$ to be appropriate. The subgroup structure was chosen to partition the patient population in treated with cfLVAD and not-treated, the object of the analysis. An initial comparison of the total wait-list time and mortality shows the benefits of cfLVAD, those patients surviving on average 256 days whereas patients without cfLVAD on average 69 days. Thus we find a clear *average* benefit of cfLVADs. Estimated `HBS` parameter values show this divergence and also markedly different patient behaviour among the two groups. The treated group displays an arc-shaped hazard $(\hat{\sigma}, \hat{\lambda}) = (1.52, -0.13)$ while the control group displays a decreasing hazard $(\hat{\sigma}, \hat{\lambda}) = (2.77, -0.54)$. These distinct behaviours could not have been learned by any of the subfamilies of the $\mathcal{GG}$ distribution, its flexibility is needed to accurately describe heterogeneous patient populations.

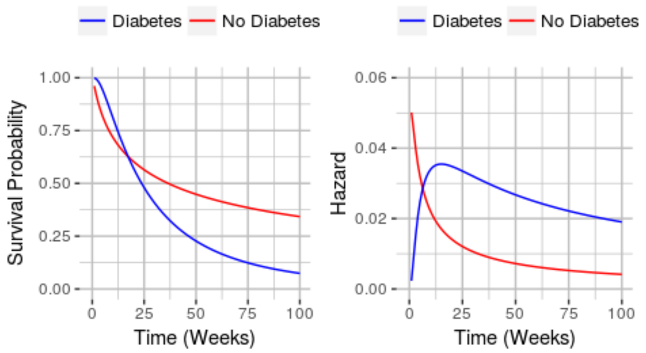

Fig. 3. Estimated average survival and hazard curves for diabetic and diabetes-free patients in the UNOS data set.

Figure 2 graphically shows the behaviour of the two groups and the comparison of the survival functions estimated by HBS with those predicted by Kaplan Meier (KM) and Cox models. The empirical KM curves for the two subgroups show crossing survival functions suggesting that cfLVADs impart an initial benefit on survival proportions (left panel). The curves estimated by HBS capture this relationship (middle panel) and moreover is able within the subgroup treated with cfLVADs to distinguish between patients. For illustrating individual estimates from HBS, the right panel shows survival curves of a 56 year-old male with no known adverse covariates and a 35 year-old female who had received a previous transplant and high doses of Creatinine. At the 100 week time horizon he has an estimated probability 40% above average survival while she has a predicted survival probability 50% below average. For illustration, on the left panel we show curves estimated by the basic Cox model, which is unable to capture the crossing behaviour.

As previously mentioned, the survival benefit of the second generation cfLVAD is heterogeneous. We applied the *personalized interpreter* to assess the impact and significance of cfLVAD on all patients in our cohort that were treated with these devices. We found that the observed effect of median survival (computed with (12)) was statistically significant ($p < 0.05$), adjusted for multiple testing with Bonferroni Correction [35], for 51% of the patients treated with cfLVAD. Moreover, a closer inspection into the feature distribution among patients for which the cfLVAD was significantly beneficial for survival and among patients for whom it was not reveals noticeable differences. Patients for which cfLVADs were not significantly beneficial had on average 3% higher weight, 4.5% more cases of diabetes, 11.5% higher levels of Creatinine and a 5% higher body mass index. Based on our findings, a more efficient cfLVAD allocation scheme can be constructed accounting for these patient characteristics, shown to be relevant for survival outcome with cfLVADs.

Analysis of co-morbidities: Diabetes

Interactions among chronic diseases in multi-morbid populations greatly affect the clinical presentation and progress of a disease of interest which has far-reaching effects on treatment design and the remedial-diagnostic process [36]. Co-morbidity has been shown to intensify health care utilization and to increase medical care costs for patients with diabetes [37]. We illustrate this interrelation with chronic cardiovascular diseases with HBS to stress the heterogeneous behaviour of multi-morbid patients as compared to patients without recorded multi-morbidity, here diabetes. Estimated average survival and hazard curves for all diabetic and diabetes-free patients are shown in Figure 3. The difference in estimated survival is striking, diabetic patients have a noticeably lower survival probability in the long term. Moreover, the shape of the survival distribution suggests a decreasing and arc-shaped hazard rate for non-diabetic and diabetic patients respectively, this is important because non-diabetic patients having survived for over 50 weeks will probably recover (hazard rate is low) while diabetic patients are still at a significant risk of an adverse event. The *personalized interpreter* applied to all patients at risk of cardiovascular diseases with diabetes systematically reveals that diabetes is associated with reducing median survival for all patients (on average by 11%), with the reduction being statistically significant for 32% of all diabetic patients in the considered cohort.

TABLE II
C-INDEX FIGURES AT SPECIFIED TIME HORIZONS SINCE WAIT-LIST REGISTRATION (HIGH C-INDEX IS BETTER).

Models	**3 Months**	**1 Year**	**3 Years**	⏱
Cox	0.628±0.029	0.611±0.022	0.589±0.024	< 1s
Weibull	0.626±0.028	0.610±0.022	0.589±0.024	< 1s
Aalen	0.625±0.028	0.610±0.022	0.591±0.023	12.0s
CoxBoost	0.632±0.033	0.614±0.028	0.593±0.028	59.5s
RSF	0.652±0.025	0.624±0.020	0.602±0.025	12.4s
CForest	0.667±0.025	0.627±0.024	0.605±0.025	1.7s
BART	0.672±0.026	0.628±0.025	0.622±0.026	424s
Wei-Tree	0.675±0.027	0.637±0.024	0.620±0.030	643s
HBS	0.697±0.025	0.646±0.021	0.627±0.026	685s

TABLE III
BRIER SCORE AT SPECIFIED TIME HORIZONS SINCE WAIT-LIST REGISTRATION (LOW BS IS BETTER).

Models	**3 Months**	**1 Year**	**3 Years**	⏱
Cox	0.206±0.010	0.237±0.004	0.245±0.010	< 1s
Weibull	0.206±0.010	0.237±0.004	0.245±0.011	< 1s
Aalen	0.206±0.011	0.238±0.005	0.245±0.010	12.0s
CoxBoost	0.207±0.008	0.238±0.003	0.242±0.010	59.5s
RSF	0.201±0.014	0.237±0.008	0.235±0.009	12.4s
CForest	0.195±0.009	0.230±0.003	0.235±0.006	1.7s
BART	0.193±0.009	0.217±0.008	0.216±0.010	424s
Wei-Tree	0.195±0.011	0.230±0.008	0.241±0.016	643s
HBS	0.191±0.009	0.225±0.008	0.225±0.012	685s

D. Predictive Performance

We consider all baseline models described in Section V-A for C-index and Brier Score comparisons at time horizons 3 months, 1 year and 3 years. All results are computed via 3 fold cross-validation with the settings of section V-C and average performance is reported. Confidence bands correspond to 95% confidence intervals across the 3 folds. The results on Table II and Table III are informative of the underlying structure in the data, we expected for instance complex feature relationships which explains the out-performance of nonparametric methods with respect to models assuming a linear covariate influence. `HBS` outperforms all baseline algorithms because we believe its flexibility, while maintaining a general parametric survival structure, is able to capture heterogeneous signals while preventing overfitting with reasonable prior restrictions (in both the survival distribution and prior parameter distributions). In contrast, purely nonparametric methods are not able to interpolate efficiently in the case of rare feature specifications and events which are prevalent in heterogeneous populations.

VII. Conclusion

Diagnostic and treatment decisions are still too often made based on the "average" patient. Understandably, the relationship of a particular treatment or condition to survival can vary greatly within populations and is important in guiding clinical decisions and understanding the disease. We introduced a flexible Bayesian ensemble method for personalized survival prediction and a model agnostic personalized interpreter to further analyze individual model outputs. The data is modelled with a rich survival distribution whose location parameter is influenced by data through a Bayesian nonparametric model and its shape and scale parameters are influenced by a subgroup partition of the population. For healthcare practitioners, this means being able to provide more personalized care to patients with greater diagnostic certainty and provide treatment plans according to the latest medical evidence in medical data. We note that we could further personalize predictions by letting covariates influence the shape and scale parameters more flexibly, this could be particularly useful if the data does not suggest a natural partition. An interesting direction for future work would be to extend this setting to competing risks outcomes similar to [38]–[40] and to inferring treatment effects [41].

References

[1] E. Abrahams and M. Silver, "The history of personalized medicine," *Integrative neuroscience and personalized medicine*, pp. 3–16, 2010.

[2] R. Snyderman, "Personalized health care: from theory to practice," *Biotechnology journal*, vol. 7, no. 8, pp. 973–979, 2012.

[3] S. D. Horn and J. Gassaway, "Practice based evidence: incorporating clinical heterogeneity and patient-reported outcomes for comparative effectiveness research," *Medical care*, vol. 48, no. 6, pp. S17–S22, 2010.

[4] S. Taghavi, S. N. Jayarajan, E. Komaroff, and A. A. Mangi, "Continuous flow left ventricular assist device technology has influenced wait times and affected donor allocation in cardiac transplantation," *The Journal of thoracic and cardiovascular surgery*, vol. 147, no. 6, pp. 1966–1971, 2014.

[5] Y. Wang, M. Simon, P. Bonde, B. U. Harris, J. J. Teuteberg, R. L. Kormos, and J. F. Antaki, "Prognosis of right ventricular failure in patients with left ventricular assist device based on decision tree with smote," *IEEE Transactions on Information Technology in Biomedicine*, vol. 16, no. 3, pp. 383–390, 2012.

[6] O. Wever-Pinzon, S. G. Drakos, A. G. Kfoury, J. N. Nativi, E. M. Gilbert, M. Everitt, R. Alharethi, K. Brunisholz, F. M. Bader, D. Y. Li *et al.*, "Morbidity and mortality in heart transplant candidates supported with mechanical circulatory support. is reappraisal of the current unos thoracic organ allocation policy justified?" *Circulation*, pp. CIRCULATIONAHA–112, 2012.

[7] D. R. Cox, "Regression models and life tables (with discussion)," *Journal of the Royal Statistical Society. Series B*, vol. 34, p. 187–220, 1972.

[8] D. W. Hosmer, S. Lemeshow, and S. May, *Parametric Regression Models*. John Wiley & Sons, Inc., 2008, pp. 244–285. [Online]. Available: http://dx.doi.org/10.1002/9780470258019.ch8

[9] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, "Random survival forests," *The annals of applied statistics*, pp. 841–860, 2008.

[10] T. Hothorn, K. Hornik, and A. Zeileis, "Unbiased recursive partitioning: A conditional inference framework," *Journal of Computational and Graphical statistics*, vol. 15, no. 3, pp. 651–674, 2006.

[11] H. A. Chipman, E. I. George, R. E. McCulloch *et al.*, "Bart: Bayesian additive regression trees," *The Annals of Applied Statistics*, vol. 4, no. 1, pp. 266–298, 2010.

[12] A. Kapelner and J. Bleich, "bartmachine: Machine learning with bayesian additive regression trees," *arXiv preprint arXiv:1312.2171*, 2013.

[13] R. A. Sparapani, B. R. Logan, R. E. McCulloch, and P. W. Laud, "Nonparametric survival analysis using bayesian additive regression trees (bart)," *Statistics in medicine*, vol. 35, no. 16, pp. 2741–2753, 2016.

[14] J. B. Nasejje, H. Mwambi, K. Dheda, and M. Lesosky, "A comparison of the conditional inference survival forest model to random survival forests based on a simulation study as well as on two applications with time-to-event data," *BMC Medical Research Methodology*, vol. 17, no. 1, p. 115, 2017.

[15] G. J. Katuwal and R. Chen, "Machine learning model interpretability for precision medicine," *arXiv preprint arXiv:1610.09045*, 2016.

[16] B. Goodman and S. Flaxman, "European union regulations on algorithmic decision-making and a" right to explanation"," *arXiv preprint arXiv:1606.08813*, 2016.

[17] T. Peltola, A. S. Havulinna, V. Salomaa, and A. Vehtari, "Hierarchical bayesian survival analysis and projective covariate selection in cardiovascular event risk prediction," in *Proceedings of the Eleventh UAI Conference on Bayesian Modeling Applications Workshop-Volume 1218*. CEUR-WS. org, 2014, pp. 79–88.

[18] V. Bonato, V. Baladandayuthapani, B. M. Broom, E. P. Sulman, K. D. Aldape, and K.-A. Do, "Bayesian ensemble methods for survival prediction in gene expression data," *Bioinformatics*, vol. 27, no. 3, pp. 359–367, 2010.

[19] F. J Diaz, H.-W. Yeh, and J. de Leon, "Role of statistical random-effects linear models in personalized medicine," *Current Pharmacogenomics and Personalized Medicine (Formerly Current Pharmacogenomics)*, vol. 10, no. 1, pp. 22–32, 2012.

[20] A. Datta, S. Sen, and Y. Zick, "Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems," in *Security and Privacy (SP), 2016 IEEE Symposium on*. IEEE, 2016, pp. 598–617.

[21] P. Adler, C. Falk, S. A. Friedler, G. Rybeck, C. Scheidegger, B. Smith, and S. Venkatasubramanian, "Auditing black-box models for indirect influence," in *Data Mining (ICDM), 2016 IEEE 16th International Conference on*. IEEE, 2016, pp. 1–10.

[22] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you?: Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 1135–1144.

[23] A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin, "Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation," *Journal of Computational and Graphical Statistics*, vol. 24, no. 1, pp. 44–65, 2015.

[24] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[25] C. Cox, H. Chu, M. F. Schneider, and A. Muñoz, "Parametric survival analysis and taxonomy of hazard functions for the generalized gamma distribution," *Statistics in medicine*, vol. 26, no. 23, pp. 4352–4374, 2007.

[26] P. Schulam, F. Wigley, and S. Saria, "Clustering longitudinal clinical marker trajectories from electronic health data: Applications to phenotyping and endotype discovery." in *AAAI*, 2015, pp. 2956–2964.

[27] W. K. Hastings, "Monte carlo sampling methods using markov chains and their applications," *Biometrika*, vol. 57, no. 1, pp. 97–109, 1970.

[28] T. Hastie and R. Tibshirani, "Bayesian backfitting (with comments and a rejoinder by the authors," *Statistical Science*, vol. 15, no. 3, pp. 196–223, 2000.

[29] T. A. Gerds, M. W. Kattan, M. Schumacher, and C. Yu, "Estimating a time-dependent concordance index for survival prediction models with

covariate dependent censoring," *Statistics in Medicine*, vol. 32, no. 13, pp. 2173–2184, 2013.

[30] U. B. Mogensen, H. Ishwaran, and T. A. Gerds, "Evaluating random forests for survival analysis using prediction error curves," *Journal of statistical software*, vol. 50, no. 11, p. 1, 2012.

[31] O. O. Aalen, "A linear regression model for the analysis of life times," *Statistics in medicine*, vol. 8, no. 8, pp. 907–925, 1989.

[32] H. Binder and M. Schumacher, "Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models," *BMC bioinformatics*, vol. 9, no. 1, p. 14, 2008.

[33] D. Meyer, J. G. Rogers, L. Edwards, E. Callahan, S. Webber, M. Johnson, J. Vega, M. Zucker, and J. Cleveland, "The future direction of the adult heart allocation system in the united states," *American Journal of Transplantation*, vol. 15, no. 1, pp. 44–54, 2015.

[34] G. J. Arnaoutakis, T. J. George, A. Kilic, C. A. Beaty, E. S. Weiss, J. V. Conte, and A. S. Shah, "Risk factors for early death in patients bridged to transplant with continuous-flow left ventricular assist devices," *The Annals of thoracic surgery*, vol. 93, no. 5, pp. 1549–1555, 2012.

[35] M. A. Napierala, "What is the bonferroni correction," *AAOS Now*, vol. 6, no. 4, p. 40, 2012.

[36] S. Mercer, C. Salisbury, and M. Fortin, *ABC of Multimorbidity*. John Wiley & Sons, 2014.

[37] V. M. van Deursen, R. Urso, C. Laroche, K. Damman, U. Dahlström, L. Tavazzi, A. P. Maggioni, and A. A. Voors, "Co-morbidities in patients with heart failure: an analysis of the european heart failure pilot survey," *European journal of heart failure*, vol. 16, no. 1, pp. 103–111, 2014.

[38] A. Bellot and M. Schaar, "Tree-based bayesian mixture model for competing risks," in *International Conference on Artificial Intelligence and Statistics*, 2018, pp. 910–918.

[39] A. M. Alaa and M. van der Schaar, "Deep multi-task gaussian processes for survival analysis with competing risks," in *Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS)*, 2017.

[40] C. Lee, W. R. Zame, J. Yoon, and M. van der Schaar, "Deephit: A deep learning approach to survival analysis with competing risks," 2018.

[41] A. M. Alaa and M. van der Schaar, "Bayesian inference of individualized treatment effects using multi-task gaussian processes," in *Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS)*, 2017.