Tree-based Bayesian Mixture Model for Competing Risks

Alexis Bellot
University of Oxford, UK

Mihaela van der Schaar
University of Oxford, UK
Alan Touring Institute, London, UK

Abstract

Many chronic diseases possess a shared biology. Therapies designed for patients at risk of multiple diseases need to account for the shared impact they may have on related diseases to ensure maximum overall well-being. Learning from data in this setting differs from classical survival analysis methods since the incidence of an event of interest may be obscured by other related competing events. We develop a semi-parametric Bayesian regression model for survival analysis with competing risks, which can be used for jointly assessing a patient's risk of multiple (competing) adverse outcomes. We construct a Hierarchical Bayesian Mixture (HBM) model to describe survival paths in which a patient's covariates influence both the estimation of the type of adverse event and the subsequent survival trajectory through Multivariate Random Forests. In addition variable importance measures, which are essential for clinical interpretability are induced naturally by our model. We aim with this setting to provide accurate individual estimates but also interpretable conclusions for use as a clinical decision support tool. We compare our method with various state-of-the-art benchmarks on both synthetic and clinical data.

1 Introduction

Life expectancy has dramatically increased in industrialized nations over the last 200 hundred years. The aging of populations carries over to clinical research and leads to an increasing representation of elderly and multimorbid individuals in study populations. Elderly individuals are likely to experience one of several disease endpoints other than the endpoint of main interest (Koller et al. 2012b). In these settings the time to occurrence of one event of interest may be obscured by other so called competing events. In fact, (Koller et al. 2012b) in a review of 50 clinical studies performed in individuals susceptible to competing risks published in high-impact clinical journals found competing risks issues in 70% of all articles. As an example, prediction of Coronary Heart Disease events in elderly subjects is known to be complicated by the fact that subjects may die from other causes prior to the observation of the disease event of interest (Wolbers et al. 2009), (Koller et al. 2012a).

Survival analysis is a method for analyzing data where the target variable is the time to the occurrence of a certain event. Competing risks is an extension where we distinguish between multiple possible events. Conceptually, we interpret a patient's overall survival path as being generated from a combination of latent trajectories related to the possible end-points/causes he/she may be at risk of (e.g. different diseases for instance may contribute to survival) even though a *primary* cause will be recorded for each patient. We model the distribution of event time $T_i \in \mathbb{R}^+$ and *primary* event cause Z_i jointly. We decompose the joint distribution into a product of latent mixing variables which we interpret as event causes, and the conditional distribution of the times given a particular cause resulting in a Mixture of distributions (Larson and Dinse 1985). Figure 1 illustrates our approach. Conventional methods for survival analysis such as the Kaplan Meier method and standard Cox proportional hazards regression ignore the dependence among competing events, and as a result underestimate true survival probabilities (see related works section) (Putter, Fiocco, and Geskus 2007). We propose a Bayesian learning approach that leverages common factors among competing events to predict parameter values supported by the data. Probabilistic statements can be made directly about the unknown model parameters and confidence on the resulting survival estimates which are needed in medical practice.

Contribution. We conceptualise the competing risks problem as a mixture of competing survival trajectories with latent variables determining the weight of these different but

Proceedings of the 21^{st} International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. JMLR: W&CP volume 7X.

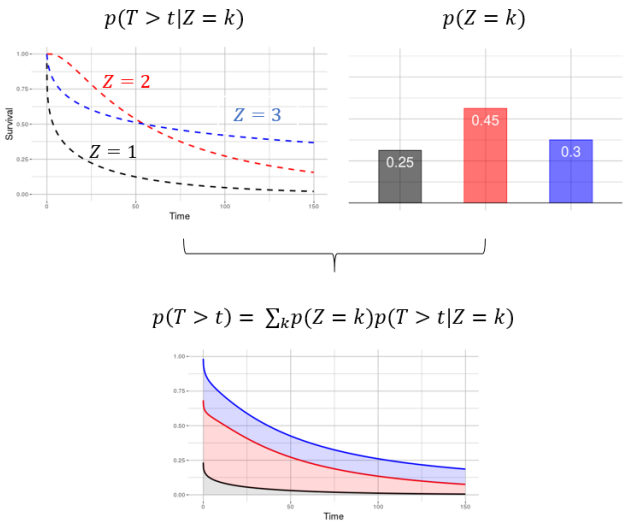

Figure 1: Depiction of the proposed approach. Overall survival function is generated from a weighted average of latent cause-dependent survival functions and probabilities of these trajectories occurring. For example, this would correspond to a patient susceptible to three diseases but with different baseline risks for each of them.

related trajectories. The parameters of the cause-specific distributions and assignment variables are modelled jointly with Multivariate Random Forest (MRF), this allows us to learn a "shared representation" of the patients survival times with respect to multiple related co-morbidities and allow for nonlinear covariate influences. The proposed model is Bayesian: we assign a prior distribution over the space of parameters, and update the posterior distribution given time-to-event data from patients at risk of competing events. This process gives rise to patient-specific survival distribution, from which a patient-specific, cause-related cumulative incidence function can be easily derived. Through the use of latent variables we naturally facilitate the incorporation of domain knowledge such as unobserved dependence hierarchies or expected prevalence of disease into the model, which enables learning clusters and groups from data. In addition, model-free variable selection and importance measures are naturally induced by our model. The hierarchical Bayesian learning framework proposed here, that leverages the interpretability of parametric distributions and predictive ability of non-parametric methods can be accommodated to incorporate other parametric distributions and regression functions to suit various applications, not necessarily in medicine.

2 Related Work

It is well understood that conventional survival models are inadequate to discriminate between competing events. Various methods have been developed for the explicit analysis of competing risk, mainly in the Statistics literature. One of the earliest attempts is due to (Prentice et al. 1978), in which they implemented standard survival models like Cox regression (Cox 1972) on a cause-specific hazard. However, covariates influence survival independently of other causes, thus not accounting for the shared structure competing events exhibit. (Fine and Gray 1999) introduced a joint regression approach focusing on the sub-distribution hazard. Their approach offers a direct interpretation in terms of survival probabilities for a particular failure type, as opposed to the cause-specific hazard model. Other assumptions such as proportional hazards and linear predictors may limit their inference in heterogeneous cohorts from modern studies even though particularly in medical application their accessible interpretation has made both of the above widely popular.

A different modelling approach was proposed by (Larson and Dinse 1985) in which competing risks are modelled as a mixture of distributions with hidden type of event assignment variables. We view our model as a Bayesian generalization of this mixture, where we use a non-linear regression function and more general event distributions.

Recently a growing interest in studying survival and competing risks is palpable also in the machine learning community. For example, Survival Random Forests (Ishwaran et al. 2008) have been adapted and applied directly to the competing risks problem in (Ishwaran et al. 2014). What differs here are the splitting rules used to grow the tree and the estimated values calculated within the terminal nodes, both based on event-specific measures. This approach – solely data-driven – gives great flexibility but often at the expense of interpretability: clinicians are unable to explain model predictions which has limited practical medical use (Lipton 2017). Many other methods have been developed for survival analysis such as deep exponential families (Ranganath et al. 2016), semi-parametric Bayesian models based on Gaussian processes (Fernández, Rivera, and Teh 2016) and deep survival neural networks (Katzman et al. 2016) but these are not directly applicable to the competing risks problem.

3 Hierarchical Bayesian Mixture

3.1 Problem Setup

In numerous medical settings we deal with a heterogeneous set of patients at risk of experiencing multiple mutually exclusive events. Each subject (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 one of the competing events occurs, which is drawn from a distribution $T_i \sim \mathbb{P}(.|X_i)$ and a categorical variable $Z_i \in \{\emptyset, 1, ..., K\}$ (with realization z_i) which indicates the type of event observed. We write $z_i = \emptyset$ for a right censored observation (i.e. a patient whose follow-up

has been interrupted) and $z_i = 1...K$ denotes one of K competing events. Figure 2 illustrates a typical competing risks scenario. As mentioned we construct a shared representation by interpreting survival as an ensemble of latent cause-specific survival paths. Thus, all estimated cause-specific survival paths potentially influence the final outcome. (Note that this interpretation differs substantially from recent machine learning methods which consider the event time to be the *minimum* of a set of cause-specific survival times and learn a shared representation through censored observations, e.g. see (Alaa and van der Schaar 2017; Lee et al. 2018)).

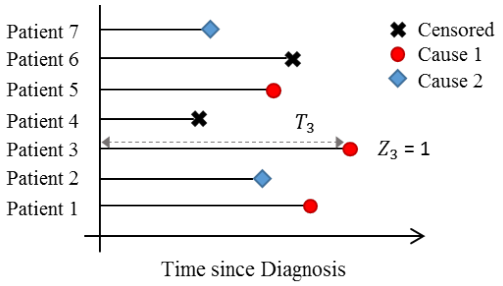

Figure 2: Illustration of survival data under competing risks.

Similar to conventional survival analysis, two identifiable quantities are of interest under competing risks: the cause-specific hazard function and the cumulative incidence function. Our goal is to estimate from an observational data set $\mathcal{D}$ that comprises n independent samples of the random tuple $\{X_i, Z_i, T_i\}$ the cumulative incidence function (CIF),

$$F(t, k|X_i) = \mathbb{P}(T_i < t, Z_i = k|X_i)$$

i.e. the probability of experiencing event k before time t. The overall distribution function is the sum of CIFs,

$$F(t|X_i) = \sum_k \mathbb{P}(T_i < t, Z_i = k|X_i)$$

The cause-specific hazard function,

$$\lambda(t, k|X_i) = \lim_{dt \to 0} \mathbb{P}(t \leq T_i \leq t + dt, Z_i = k| \quad T_i \geq t, X_i)/dt \tag{1}$$

represents the instantaneous risk of experiencing an end-point related to cause k and indicates the rate at which mortality with respect to that cause progresses with time. A similar expression can be derived for the overall hazard.

3.2 Model

The conceptual structure is that of a generative probabilistic mixture model constructed in a hierarchical fashion. We compute patient-specific survival estimates by modelling the survival time T_i directly as a function of the patients covariates through a generative probabilistic mixture model. We decompose the joint distribution (T_i, Z_i) as $\mathbb{P}(T_i, Z_i) = \mathbb{P}(T_i|Z_i)\mathbb{P}(Z_i)$ and write,

$$T_i|Z_i = k \sim \mathcal{GG}(\beta_{ik}, \sigma_i, \lambda_i) \tag{2}$$

$$Z_i \sim Cat(\pi_{i1}, ..., \pi_{iK}) \tag{3}$$

The time until an end-point related to cause k, $T_i|Z_i = k$, $i = 1...n$ is assumed to be generated from a Generalized Gamma distribution $(\mathcal{GG})$ (Cox et al. 2007). The motivation is that it contains as special cases most of the familiar distributions used in survival settings such as the Weibull $(\lambda = 1)$, Gamma $(\sigma = \lambda)$ and Log-Normal $(\lambda = 0)$ distributions but also its parameters relate to meaningful medical quantities such as the hazard shape (which is unavailable in 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 survival 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.

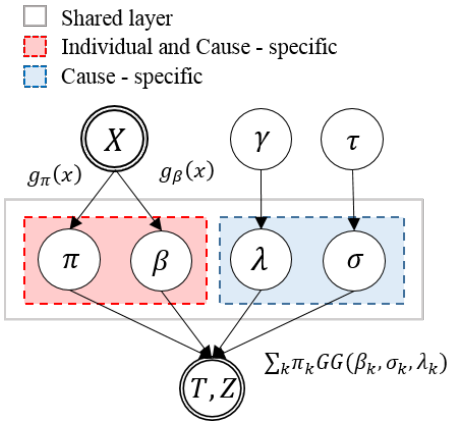

Figure 3: Graphical model induced by the HBM model. Observable variables are in double-circled nodes.

For patients in a competing risk setting we assume survival time to be generated from a Bayesian mixture of $\mathcal{GG}$ distributions. The density d of patient i is thus defined as:

$$d(t; \boldsymbol{x}_i) := \sum_{k=1}^{K} \pi_{ik} f(t; \beta_{ik}, \sigma_k, \lambda_k), \qquad t > 0 \tag{4}$$

Prior distributions on all latent parameters are introduced to exploit the hierarchical process from which the observed data is assumed to be generated. The graphical model in Figure 3 illustrates this hierarchical structure. Motivated by ensemble methods we leverage the relationship between the functional terms, each optimizing a specific aspect of the survival distribution. For instance, the distributional shape is cause-specific while the median survival and baseline risk for a disease are unique to each individual.

Mixture Regression model. The latent variables enable a shared representation to explicitly model the influence of competing events. It consists of the variables $\boldsymbol{\beta}_i = (\beta_{i1}, ..., \beta_{iK})$ and $\boldsymbol{\pi}_i = (\pi_{i1}, ..., \pi_{iK})$ which propagate the influence of covariates through a multivariate regression model. In particular $T_i, Z_i \perp\!\!\!\perp X_i | \boldsymbol{\beta}_i, \boldsymbol{\pi}_i$. Let $g_\beta, g_\pi : \mathbb{R}^d \to \mathbb{R}^K$ be multivariate regression functions, we write generally,

$$\boldsymbol{\beta}_i | \boldsymbol{x}_i \sim g_\beta(\boldsymbol{x}_i) + \epsilon_{\beta,i}, \quad \epsilon_{\beta,i} \sim \mathcal{N}(0, \alpha_\beta^2) \tag{5}$$

$$\boldsymbol{\pi}_i | \boldsymbol{x}_i \sim l(g_\pi(\boldsymbol{x}_i) + \epsilon_{\pi,i}), \quad \epsilon_{\pi,i} \sim \mathcal{N}(0, \alpha_\pi^2) \tag{6}$$

where $l(x)_i = x_i / \sum_i x_i,\ x_i > 0, \forall i$ and $(\alpha_\beta^2, \alpha_\pi^2)$ are fixed hyper-parameters. To accommodate for the wide variability in individual features and their impact on observed survival dynamics, we model g with a Multivariate Random Forest. Multivariate random trees and forests (Segal 1992; Segal and Xiao 2011) are extensions to the regression tree framework described in (Breiman 2001). Following (Segal and Xiao 2011), the empirical covariance matrix is used as part of a node impurity measure based on the mean squared error to determine homogeneous children nodes. The prediction for each leaf of a constituent regression tree is the vector of mean values for covariates reaching that leaf.

Our approach departs from other tree-based methods because we adopt a Bayesian setting. We repeatedly sample tree structures to approximate the posterior distribution. We interpret this process as exploring different multi-output tree configurations supported by the data, described probabilistically in terms of a posterior distribution. The use of Multivariate Random Forest allows us to *jointly* represent complex interactions with covariates without the need to assume a predefined non-linear transformation on the covariate space as it is the case in standard linear regression. Tree based methods are appealing in medical contexts since they have the advantage of providing prognosis based on multiple features without prior selection and are robust (not affected by monotonic transformations such as scaling or shifting of the data) to measurement errors and outliers often present in medical data. We note however that other regression function choices can be easily incorporated.

Heterogeneous cohorts at different levels of risk to various death causes might behave very differently within a population. The hazard shape of each cause-specific survival distribution in particular may exhibit different forms. For instance, cancer patients undergoing chemotherapy may see higher risk of death in the short term in contrast with the longer term than patients with higher cardiovascular disease risk. 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 without prior specifications by using the rich distributional family $\mathcal{GG}$ (Cox et al. 2007). Let $\mathcal{C}_j, j = 1, ..., K$ denote the index set of patients experiencing event cause j. We model,

$$\sigma_i \sim \sum_{j=1}^{K} \mathbb{1}\{i \in \mathcal{C}_j\} \tau_j, \quad \tau_j \sim \mathcal{G}(\eta_0, \eta_1) \tag{7}$$

$$\lambda_i \sim \sum_{j=1}^{K} \mathbb{1}\{i \in \mathcal{C}_j\} \gamma_j, \quad \gamma_j \sim \mathcal{N}(\gamma_0, \gamma_1) \tag{8}$$

$\mathcal{G}$ denotes the Gamma distribution and $(\eta_0, \eta_1, \gamma_0, \gamma_1)$ are fixed hyper-parameters. Prior distributions will be set in practice with previous domain knowledge, if available, to encourage a known survival behaviour. For a domain agnostic approach we propose choosing prior distributions by sampling to be weakly informative about the survival shapes, that is prior samples generate plausible and general survival curves not restricting posterior inference.

3.3 Learning and Inference

In the Bayesian paradigm learning parameter values and predicting hidden variables $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\sigma}, \boldsymbol{\lambda}, \boldsymbol{\pi})$ rests on computing the posterior distribution given the data and model, which provides uncertainty estimates and parameter values minimizing a variety of Bayesian loss functions. The posterior is given by 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 3 the joint posterior is given by,

$$p(\boldsymbol{\beta}, \boldsymbol{\sigma}, \boldsymbol{\lambda}, \boldsymbol{\pi} | \mathcal{D}) \propto p(\mathcal{D} | \boldsymbol{\beta}, \boldsymbol{\sigma}, \boldsymbol{\lambda}, \boldsymbol{\pi}) \times \prod_i p(\boldsymbol{\beta}_i) p(\boldsymbol{\pi}_i) p(\sigma_i) p(\lambda_i) \tag{9}$$

Let $f_k(t_i; \theta_i)$ denote the *pdf* of the random variable $T_i | Z_i = k$ and $\mathbb{P}(Z_i = k) = \pi_{ik}$. The contribution to the likelihood of an individual with end-point t_i of cause k is $\pi_{ik} f_k(t_i; \boldsymbol{\theta})$ while a censored observation j contributes $S(t_j; \boldsymbol{\theta}) = 1 - F(t_j; \boldsymbol{\theta})$. The likelihood of the observed data is given by,

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_i \sum_k \left(\pi_{ik} f_k(t_i; \boldsymbol{\theta})\right)^{\mathbb{1}\{z_i = k\}} S(t_i; \boldsymbol{\theta})^{\mathbb{1}\{z_i = 0\}} \tag{10}$$

The expressions involved make direct posterior inference intractable. To approximate the posterior we rely on sampling from a Markov Chain with target distribution the posterior in (9) using an adaptive Metropolis within Gibbs Markov Chain Monte Carlo (MCMC) scheme (Hastings 1970). We update the distributional parameters and latent variables by cycling through the parameter space updating each component sequentially. In each iteration the prior means of the latent variables $\boldsymbol{\beta}$ and $\boldsymbol{\pi}$ are updated with the multi-output tree structure (each sequence $\boldsymbol{\beta}$ and $\boldsymbol{\pi}$ jointly) that inform the prior likelihood of these parameters, thus through these latent variables a patient's covariates indirectly influences posterior parameter estimates. Next we update the distributional parameters $(\boldsymbol{\beta}, \boldsymbol{\pi}, \boldsymbol{\sigma}, \boldsymbol{\lambda})$ by cycling through the parameter space sequentially with a metropolis step (tractable since the likelihood and priors are fully specified). The proposal distribution for sampling new states in the markov chain is taken to be a gaussian distribution with an adaptive step size (its variance) updated every 50 iterations to ensure an acceptance rate of around 40%. Algorithm 1 details the complete procedure.

3.4 Posterior Variable Importance Learning

We interpret variable importance as a stochastic quantity related the latent variables $\boldsymbol{\beta}$ and $\boldsymbol{\pi}$. In the context of competing risks, within our learning algorithm we are able to differentiate between the variables that are influential in determining the absolute risk of a specific end-point e.g. probability of death due to CVD as opposed to Cancer, and the influential variables that determine the latent survival trajectory for a specific cause through g_β and g_π. For this, we leverage the model-free variable importance summaries provided by tree-based algorithms introduced in (Ishwaran and others 2007). They proposed a permutation-based approach, as the difference between normalized prediction error (mean squared error) when the variable of interest is randomly permuted versus the normalized prediction error otherwise. Let $e^*_{j,\beta}$ and $e^*_{j,\pi}$ denote the mean squared error of models $\hat{g}_\beta$ and $\hat{g}_\beta$ over the training data with variable j randomly shuffled. Then define the importance of variable j, v_j, as,

$$v_{k,j} := |e_k - e^*_{j,k}|, \quad k = \beta, \pi \tag{11}$$

where $e_k, k = \beta, \pi$ denotes the mean squared error without shuffling. The intuition is that variables that significantly alter individual predictions will have been used as splitting rules in many tree configurations suggesting high predictive power relative to other variables. Our model induces a Bayesian, probabilistic variable importance distribution explored by the MCMC sampler, each iteration leading to a different tree configuration and thus associations and variable importance summaries. The variable configurations that are strongly supported by the data may appear in most of the MCMC samples, while others with less evidence may appear less often. This approach accounts for the uncertainty in the data and gives a measure of variable importance for each one of the event causes considered and also differentiates the variables that influence the mixing components versus the mixing distributions that form the mixture model. This process gives rise to a distribution of variable importance for each covariate from which we can compute the probability of *no* effect (0 error or less) which in turn relates to the probability of a false positive if the variable being considered were called a discovery or significant. Thus, we can interpret this value as a Bayesian q-value, as in (Storey and others 2003). We could then proceed by controlling for a desired global false discovery rate (FDR) bound and determine the significance threshold controlling for the expected global Bayesian FDR as discussed in Section 4 of (Morris et al. 2008).

Algorithm 1: HBM Learning

Input: Dataset $\mathcal{D}$, number of iterations T.
Set prior distributions for $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\sigma}, \boldsymbol{\lambda}, \boldsymbol{\pi})$;
Initialize $\boldsymbol{\theta}^{(0)} = (\boldsymbol{\beta}^{(0)}, \boldsymbol{\sigma}^{(0)}, \boldsymbol{\lambda}^{(0)}, \boldsymbol{\pi}^{(0)})$;
for *t from 1 to T* **do**

- Learn $g_\pi^{(t-1)} : \mathcal{X} \to \boldsymbol{\pi}^{(t-1)}$ and $g_\beta^{(t-1)} : \mathcal{X} \to \boldsymbol{\beta}^{(t-1)}$;
- Compute variable importance $\boldsymbol{v}_\beta^{(t)} := (v_{\beta,j})_j$ and $\boldsymbol{v}_\pi^{(t)} := (v_{\pi,j})_j$ with (11);
- Update prior means $\mathbb{E}\left(\boldsymbol{\pi}^{(t)}\right) := g_\pi^{(t-1)}(\boldsymbol{X})$ and $\mathbb{E}\left(\boldsymbol{\beta}^{(t)}\right) := g_\beta^{(t-1)}(\boldsymbol{X})$;
- **for** *i from 1 to N* **do**
  $\boldsymbol{\beta}_i^{(t)} \leftarrow$ sample from Markov chain with target $p(\boldsymbol{\beta}_i | \boldsymbol{\beta}_{-i}^{(t-1)}, \boldsymbol{\sigma}^{(t-1)}, \boldsymbol{\lambda}^{(t-1)}, \boldsymbol{\pi}^{(t-1)}, \mathcal{D})$;
  end
- **for** *i from 1 to N* **do**
  $\boldsymbol{\pi}_i^{(t)} \leftarrow$ sample from Markov chain with target $p(\boldsymbol{\pi}_i | \boldsymbol{\pi}_{-i}^{(t-1)}, \boldsymbol{\sigma}^{(t-1)}, \boldsymbol{\lambda}^{(t-1)}, \boldsymbol{\beta}^{(t)}, \mathcal{D})$;
  end
- $\boldsymbol{\sigma}^{(t)} \leftarrow$ sample from Markov chain with target $p(\boldsymbol{\sigma} | \boldsymbol{\beta}^{(t)}, \boldsymbol{\sigma}^{(t-1)}, \boldsymbol{\lambda}^{(t-1)}, \boldsymbol{\pi}^{(t)}, \mathcal{D})$;
- $\boldsymbol{\lambda}^{(t)} \leftarrow$ sample from Markov chain with target $p(\boldsymbol{\lambda} | \boldsymbol{\beta}^{(t)}, \boldsymbol{\sigma}^{(t)}, \boldsymbol{\lambda}^{(t-1)}, \boldsymbol{\pi}^{(t)}, \mathcal{D})$;

end
Output: Approximate samples $\left(\boldsymbol{\theta}^{(t)}\right)_{t=1}^T$ from $p(\boldsymbol{\theta}|\mathcal{D})$ and Variable Importance samples $\left(\boldsymbol{v}^{(t)}\right)_{t=1}^T$

Given the elaborate nature of medical survival dynamics, it is of clinical interest to explore the interactions and effects of covariates on survival. This method discovers influential

variables, quantifies the uncertainty around its estimates via credible intervals and controls for false discovery rates.

4 Experiments

The purpose of the model is to provide accurate individualized predictions but also provide a description of survival dynamics. We validate our model by conducting a set of experiments on synthetic and observational data.

4.1 Performance Assessment

Due to the presence of censoring and competing risks in survival data, traditional performance metrics need to be accommodated to account for this partial information. In this paper we adopt a common approach used in the literature: the cause-specific concordance index (C-index). Formally, we define the (time-dependent) concordance index (C-index) for a cause k as follows (Wolbers et al. 2014):

$$C_k(t) := \mathbb{P}(\hat{F}_k(t; X_i) > \hat{F}_k(t; X_j) | \{z_i = k\} \wedge \{T_i \leq t\} \wedge \{T_i < T_j \vee \delta_j \neq k\}) \quad (12)$$

where $F_k(t; X_i) := \mathbb{P}(T_i < t | Z_i = k, X_i)$ is the cause-specific *cdf*. The time-dependent C-index as defined above corresponds to the probability that predicted cause-specific survival probabilities are ranked in *accordance* to the actual observed survival times given the occurrence of an event and corresponding cause. The C-index thus serves as a measure of the discriminative power for a cause of interest of a model. The measure is bound on the interval $[0.5, 1]$ in which random guessing corresponds to a C-index of 0.5 and perfect prediction to a C-index of 1. In all experiments the C-index is adjusted for censoring using inverse probability of censoring weights and approximated with the implementations in the R package `pec`.

Benchmarks. We compare our model with three baseline algorithms specifically modelling survival data under competing risks. We consider first the Cox proportional hazards model (CPH) (Cox 1972) studied in (Austin, Lee, and Fine 2016) by modelling directly the cause-specific hazard function shown in equation (1). As a second baseline the Fine-Gray proportional subdistribution hazards model (FG) which also imposes proportional hazards was introduced in (Fine and Gray 1999) but differs from CPH by modelling the sub-distribution hazard and thus enables a direct interpretation in terms of cause-specific survival probabilities, unavailable in CPH. These two baselines encode a linear effects of covariates on survival. As a nonparametric alternative we consider Competing Risks Random Forests (CRF) introduced in (Ishwaran et al. 2014), which mimic the construction of Random Forests adapting its splitting rules and leaf node predictions to cause-specific survival outcomes. Both CPH and FG do not require hyperparameter tuning and are implemented off the shelf with the R package `riskRegression`. For CRF, we followed the recommended hyper-parameter settings in (Ishwaran et al. 2014), the forest was grown with 1000 trees using a modified weighted log-rank splitting rule modelled after Grays test and minimum terminal node size was set to 6.

4.2 Synthetic data

We demonstrate the ability of our model to cope with heterogeneous populations by evaluating our model on a synthetic model with different cause-dependent interactions between survival times and covariates. We consider two scenarios with two competing events – $Z_i \in \{\emptyset, 1, 2\}$.

Scenario 1

$$X_i \sim \mathcal{U}(-2, 2)$$
$$T_i^1 \sim X_i^3 + \mathcal{N}(15, 1)$$
$$T_i^2 \sim 5X_i^2 + \mathcal{N}(5, 1)$$
$$T_i := \begin{cases} T_i^1, \text{ w. prob. } 0.8 \cdot I + 0.2 \cdot (1 - I) \\ T_i^2, \text{ w. prob. } 0.2 \cdot I + 0.8 \cdot (1 - I) \end{cases}$$

Scenario 2

$$\boldsymbol{X}_i \sim \mathcal{U}(0, 1)$$
$$T_i^1 \sim log\mathcal{N}\left(4 + \cosh(\gamma_1^T \boldsymbol{X}_i), 2\right)$$
$$T_i^2 \sim \mathcal{W}(1/2, 4 + \gamma_2^T \boldsymbol{X}_i)$$
$$T_i := \begin{cases} T_i^1, \text{ w. prob. } \Phi(\gamma_3^T \boldsymbol{X}_i) \\ T_i^2, \text{ w. prob. } 1 - \Phi(\gamma_3^T \boldsymbol{X}_i) \end{cases}$$

Where $I := \mathbb{1}\{X_i \in (-1, 1)\}$ and $\mathbb{1}$ is the indicator function. We assume two very different scenarios to illustrate our model. Scenario 1 posits quadratic and cubic covariate effects on survival for the two causes drawn from a normal distribution which a priori is unfavourable to our model which has an asymmetric form. We draw a dataset with 500 observations from this scenario. The experiment is designed to favour cause 1 events for generated event times with covariate values in the interval $(-1, 1)$ and favour cause 2 otherwise. These are shown as red and blue dots respectively on Figure 4, the solid lines are the median cause-specific survival time estimated by our model and the shaded areas are 95% credible intervals. Overall the median estimates of cause-specific survival capture the non-linear relationships for the two causes but accuracy is impacted by the large imbalance of events in the covariate distribution which results in wider credible intervals. In these cases we believe that HBM learns from the inbuilt shared representation to provide more conservative estimates closer to observations from other causes, as appears to be the case in the lower spectrum of X.

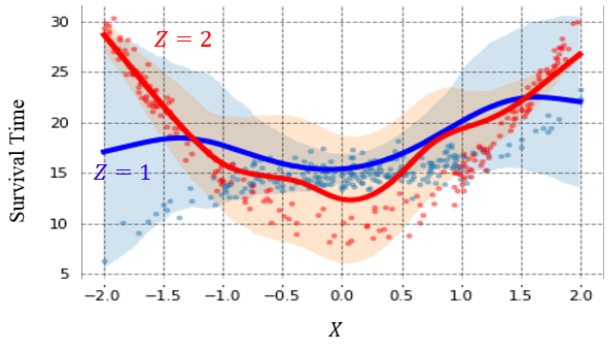

Figure 4: Generated survival times (dots) and estimated median survival times (curves) plotted against the covariate X. Blue and red colored objects correspond to observations from cause 1 and 2 respectively. Shaded areas are 95% credible intervals.

Scenario 2 simulates a more realistic heterogeneous patient population in which an end-point due to cause 1 is drawn from a log-normal distribution and an end-point due to cause 2 from a Weibull distribution (two common survival distributions). The parameters γ_1, γ_2 and γ_3 are 4 dimensional vectors we fix to generate reasonable event times. We draw 10 data sets $\mathcal{D}$ of 500 instances from scenario 2 setting aside 250 of each of them for testing, reported performance estimates are averaged over all data sets. We induce censoring by randomly selecting 20% of observed events, and for each of those set $C_i \leftarrow \mathcal{U}(0, T_i)$, their censoring time. As we can see in Table 1 HBM outperforms all other algorithms. This is because scenario 2 displays a highly nonlinear relationships between covariates and survival times, and in addition, it assumes different forms for the cause-specific distributions of the survival times, all of which are features that can be captured well by HBM but not by the other benchmarks.

Algorithms	C_1	C_2
CPH	0.575 ± 0.01	0.573 ± 0.01
FG	0.564 ± 0.01	0.581 ± 0.01
CRF	0.580 ± 0.01	0.593 ± 0.01
HBM	0.637 ± 0.01	0.666 ± 0.01

Table 1: Cause-specific $C-$index at the last observed time on the testing set of Scenario 2. Uncertainty bands are standard deviations.

4.3 SEER data

Cardiovascular disease (CVD) and breast cancer are the largest contributors to the burden of chronic disease in the United States (Hoyert, Xu, and others 2012). There is increasing evidence of overlap in risk factors and disease prevention for CVD and breast cancer suggesting that these seemingly diverse diseases have some common biological traits (Koene et al. 2016). Moreover, breast cancer treatments are suspected to accelerate or worsen pre-existing cardiac disease since both chemotherapy and radiation causes long term cardiovascular side effects. Overall mortality for these patients cannot be assessed without joint prognosis of both CVD and cancer related risk.

We investigate a patient population extracted from the Surveillance, Epidemiology, and End Results (SEER) Program. SEER is a public database [1] which provides information on cancer statistics in an effort to reduce the cancer burden among the U.S. population. The extracted cohort comprises 1000 patients described by 12 covariates including: age, gender, tumor size and type, morphology information, surgery information and a number of physiological markers related to cancer. Overall mortality was 28.2% divided into death due to CVD (2%), Cancer (17.5%) and Other (8.7%).

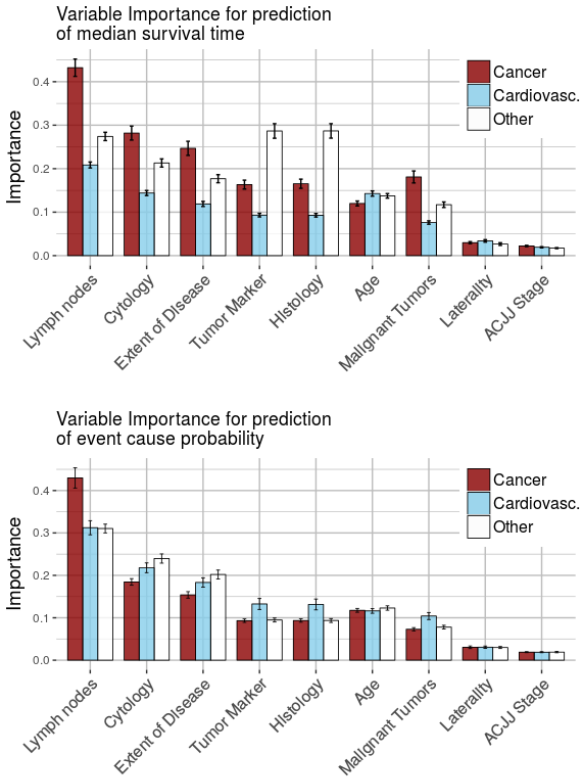

Figure 5: Variable importance for selected covariates from the SEER data set. The upper panel shows variable importance with respect to median survival while the lower panel shows variable importance with respect to baseline risk for the various causes. Confidence around mean estimates are 90% credible intervals.

[1] Available at `https://seer.cancer.gov/`

Algorithms	Cancer	Cardiovasc.	Other
CPH	0.630 ± 0.02	0.602 ± 0.08	0.647 ± 0.07
FG	0.627 ± 0.01	0.595 ± 0.06	0.632 ± 0.06
CRF	0.690 ± 0.04	0.651 ± 0.09	0.683 ± 0.04
HBM	0.721 ± 0.02	0.634 ± 0.07	0.711 ± 0.08

Table 2: $C-$index for all algorithms on the extracted dataset from SEER. Confidence bands are standard deviations.

It is of clinical interest to discover the strength of relationships between conditions or factors and survival to guide treatment design and further our understanding of the disease. The described variable importance method distinguishes between factors influential in different causes but also factors influential in predicting the marginal risk of experiencing cause k, $\mathbb{P}(Z = k)$ in contrast to factors influential in predicting median survival conditional on experiencing a particular cause k, median$(T|Z = k)$. Figure 5 shows our findings for the variables and patients extracted from the SEER data. We notice for instance that for median cause-specific survival (the upper panel), most variables recorded are more predictive of breast cancer than of cardiovascular diseases. Less divergence in variable predictive power is observed for event probability prediction (lower panel) which suggests that at onset, no recorded variable (besides "Lymph nodes") has more influence in determining baseline risk for one disease versus another. We applied the proposed variable selection procedure controlling for a global FDR bound of 5%, we find that only the first 5 variables namely, number of lymph nodes, cytology, the extent of disease, tumor marker and the number of malignant tumors significantly impact cancer outcomes. In turn for Cardiocavscular disease outcomes the 4 variables: number of lymph nodes, cytology, the extent of disease and tumor marker have a significant effect (for both survival and event cause prediction).

Table 2 provides performance estimates for all algorithms and causes at time horizon 7 years. We computed the $C-$index by 3 fold cross-validation. As was observed in Figure 5 most covariates are predictive of mortality due to breast cancer and "other causes" in the SEER dataset, which we believe explains the under-performance of all models in predicting death related to cardiovascular diseases. We note also the large confidence intervals in all estimates, the SEER data has very low mortality and thus variability is expected since censored observations contribute only indirectly to the C-index. HBM provides competitive performance in average estimates compared to CRF and substantial improvements with respect to CPH and FG which suggests that a shared representation with a nonlinear predictor is helpful in explaining the complex nature of competing risks.

5 Conclusion

Competing risks settings are complex and interlaced, it happens that most real world medical problems are of this type. To improve clinical practice in the prognosis and treatment of complex diseases, and discover what factors and how they affect different diseases, it is crucial to account for heterogeneous cohorts and shared influences of covariates on survival from a specific cause. We have proposed a Bayesian model specifying these relationships in different aspects of the overall survival path to provide an intuitive representation. We provide confidence estimates and assess the importance of variables for each cause. Although the methods presented here represent only some steps along the way, they yield quantitative and qualitative improvements over previous methods.
From a medical perspective our model contributes towards the field of precision medicine, there is growing awareness among clinicians that to improve the response to therapy and long term prognosis, treatment must be specifically tailored to the disease and the patient. Based on overall as well as cause-specific individual survival estimates provided by our model clinicians can optimize treatment allocation schemes and more accurately weight the benefits of a treatment for a particular disease which may have side-effects on the risk of related diseases. Through the personalized estimates of survival and variable importance offered by our model, we hope clinicians can further their understanding of connected diseases and improve health care delivery.

References

Alaa, A., and van der Schaar, M. 2017. Deep multi-task gaussian processes for survival analysis with competing risks. *NIPS*.

Austin, P. C.; Lee, D. S.; and Fine, J. P. 2016. Introduction to the analysis of survival data in the presence of competing risks. *Circulation* 133(6):601–609.

Breiman, L. 2001. Random forests. *Machine learning* 45(1):5–32.

Cox, C.; Chu, H.; Schneider, M. F.; and Muñoz, A. 2007. Parametric survival analysis and taxonomy of hazard functions for the generalized gamma distribution. *Statistics in medicine* 26(23):4352–4374.

Cox, D. R. 1972. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society. Series B* 34:187220.

Fernández, T.; Rivera, N.; and Teh, Y. W. 2016. Gaussian processes for survival analysis. In *Advances in Neural Information Processing Systems*, 5021–5029.

Fine, J. P., and Gray, R. J. 1999. A proportional hazards model for the subdistribution of a competing risk. *Journal of the American statistical association* 94(446):496–509.

Hastings, W. K. 1970. Monte carlo sampling methods using markov chains and their applications. *Biometrika* 57(1):97–109.

Hoyert, D. L.; Xu, J.; et al. 2012. Deaths: preliminary data for 2011. *Natl Vital Stat Rep* 61(6):1–51.

Ishwaran, H., et al. 2007. Variable importance in binary regression trees and forests. *Electronic Journal of Statistics* 1:519–537.

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

Ishwaran, H.; Gerds, T. A.; Kogalur, U. B.; Moore, R. D.; Gange, S. J.; and Lau, B. M. 2014. Random survival forests for competing risks. *Biostatistics* 15(4):757–773.

Katzman, J.; Shaham, U.; Bates, J.; Cloninger, A.; Jiang, T.; and Kluger, Y. 2016. Deep survival: A deep cox proportional hazards network. *arXiv preprint arXiv:1606.00931*.

Koene, R. J.; Prizment, A. E.; Blaes, A.; and Konety, S. H. 2016. Shared risk factors in cardiovascular disease and cancer. *Circulation* 133(11):1104–1114.

Koller, M. T.; Leening, M. J.; Wolbers, M.; Steyerberg, E. W.; Hunink, M. M.; Schoop, R.; Hofman, A.; Bucher, H. C.; Psaty, B. M.; Lloyd-Jones, D. M.; et al. 2012a. Development and validation of a coronary risk prediction model for older us and european persons in the cardiovascular health study and the rotterdam study. *Annals of internal medicine* 157(6):389–397.

Koller, M. T.; Raatz, H.; Steyerberg, E. W.; and Wolbers, M. 2012b. Competing risks and the clinical community: irrelevance or ignorance? *Statistics in Medicine* 31(11-12):1089–1097.

Larson, M. G., and Dinse, G. E. 1985. A mixture model for the regression analysis of competing risks data. *Applied statistics* 201–211.

Lee, C.; Zame, W. R.; Yoon, J.; and van der Schaar, M. 2018. Deephit: A deep learning approach to survival analysis with competing risks. *AAAI*.

Lipton, Z. C. 2017. The doctor just won't accept that! *arXiv preprint arXiv:1711.08037*.

Morris, J. S.; Brown, P. J.; Herrick, R. C.; Baggerly, K. A.; and Coombes, K. R. 2008. Bayesian analysis of mass spectrometry proteomics data using wavelet based functional mixed models.

Prentice, R. L.; Kalbfleisch, J. D.; Peterson Jr, A. V.; Flournoy, N.; Farewell, V. T.; and Breslow, N. E. 1978. The analysis of failure times in the presence of competing risks. *Biometrics* 541–554.

Putter, H.; Fiocco, M.; and Geskus, R. B. 2007. Tutorial in biostatistics: competing risks and multi-state models. *Statistics in medicine* 26(11):2389–2430.

Ranganath, R.; Perotte, A.; Elhadad, N.; and Blei, D. 2016. Deep survival analysis. In *Machine Learning for Healthcare Conference*, 101–114.

Segal, M., and Xiao, Y. 2011. Multivariate random forests. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1(1):80–87.

Segal, M. R. 1992. Tree-structured methods for longitudinal data. *Journal of the American Statistical Association* 87(418):407–418.

Storey, J. D., et al. 2003. The positive false discovery rate: a bayesian interpretation and the q-value. *The Annals of Statistics* 31(6):2013–2035.

Wolbers, M.; Koller, M. T.; Witteman, J. C.; and Steyerberg, E. W. 2009. Prognostic models with competing risks: methods and application to coronary risk prediction. *Epidemiology* 20(4):555–561.

Wolbers, M.; Blanche, P.; Koller, M. T.; Witteman, J. C.; and Gerds, T. A. 2014. Concordance for prognostic models with competing risks. *Biostatistics* 15(3):526–539.