Boosted Trees for Risk Prognosis

Alexis Bellot ALEXIS.BELLOT@ENG.OX.AC.UK
Department of Engineering Science
University of Oxford
Oxford, United Kingdom

Mihaela van der Schaar MIHAELA.VANDERSCHAAR@ENG.OX.AC.UK
Department of Engineering Science
University of Oxford
Oxford, United Kingdom

Abstract

We present a new approach to ensemble learning for risk prognosis in heterogeneous medical populations. Our aim is to improve overall prognosis by focusing on under-represented patient subgroups with an atypical disease presentation; with current prognostic tools, these subgroups are being consistently mis-estimated. Our method proceeds sequentially by learning nonparametric survival estimators which iteratively learn to improve predictions of previously misdiagnosed patients - a process called *boosting*. This results in fully nonparametric survival estimates, that is, constrained neither by assumptions regarding the baseline hazard nor assumptions regarding the underlying covariate interactions - and thus differentiating our approach from existing boosting methods for survival analysis. In addition, our approach yields a measure of the relative covariate importance that accurately identifies relevant covariates within complex survival dynamics, thereby informing further medical understanding of disease interactions. We study the properties of our approach on a variety of heterogeneous medical datasets, demonstrating significant performance improvements over existing survival and ensemble methods.

1. Introduction

Boosting (Freund and Schapire, 1995; Freund et al., 1999) is a general ensemble machine learning approach that combines simple predictive models (also referred to as hypotheses) trained sequentially such that each one of them is explicitly encouraged to correct mistakes of previous hypotheses. We consider the problem of predicting event probabilities over time, such as death or onset of disease, with the aim of providing a *fully individualized survival function* for each patient. This effectively extends the general approach of boosting to event-time estimation, a setting which differs from the more familiar classification and regression problems.

The intuition of learning from the performance of previous hypotheses is especially appealing for improving predictions for patients that are consistently being mistreated with current policies. (Skinner et al., 2016; Falchuk and Falchuk, 2012) show that misdiagnoses in patients with atypical disease presentation or risk factors represent a major source of patient harm. Their missed prognoses can be attributed in part to predictive models that do not accurately capture the heterogeneous patterns of disease present in modern patient populations, often because predictive models lack the flexibility to provide truly personalized predictions. A revealing example relates to patient populations at risk of cardiovascular diseases (CVD), associated with large health and economic burdens worldwide (Benjamin et al., 2017). In this case much of this burden is due to missed or delayed diagnoses in patients with no known risk factors or unusual symptoms (Quinn et al., 2017). This is the result of the excessively complex nature of the disease and its interactions with risk factors for which the underlying causal biological traits are poorly understood (Yao et al., 2014; MacLellan et al., 2012). For instance, (Kathiresan and Srivastava, 2012) show that even within a narrow phenotype, mortality rates can be highly divergent. Despite these findings, currently used medical risk scores are composed of linear associations of only few known risk factors (Wong et al., 2014; Schnabel et al., 2009) which do not accurately discriminate between patients.

Our approach intends to precisely focus on complex patterns and subgroups of patients that are consistently being misdiagnosed. Our goal is to use this intuition for improving risk prognosis without imposing assumptions such as proportional hazards - which restricts the rate of mortality of two populations to be in constant proportion over time-, accelerated failure times or pre-specified interactions between covariates and survival. We develop two nonparametric boosting-based algorithms that iteratively train shallow survival trees on samples of the patient population. After each iteration, the patient population is re-weighted so as to bias the next iteration toward correcting previous errors in the predicted survival function. Final survival estimates result from a weighted average of individual tree predictions dependent on each tree's predictive performance. Our method provides an efficient scheme for learning in high-dimensional settings at a low computational cost and hence, it is able to leverage the full breadth of large medical health records.

We supplement our prognostic model with a post-processing step to assess covariate influence in determining survival predictions which can guide clinicians in better understanding the biology of the disease, particularly when a priori knowledge is scarce. We do this by examining the improvement in a measure of goodness of fit due to a particular risk factor - e.g. diabetes, cholesterol level etc.

Technical Significance We develop an extension of boosting architectures to survival analysis, the problem of predicting the occurrence of events in time. In contrast to single time predictions (such as classification or regression), we estimate *full* probability distributions and propose a notion of prediction "correctness" which successfully drives weak learners towards frequently mis-estimated patients over successive iterations.

Clinical Relevance From a medical perspective our model contributes towards the field of "precision medicine". Patient heterogeneity is one of the major reasons for the large share of misdiagnoses in chronic diseases. We present a predictive model designed to leverage the heterogeneity present in large modern data sets -by precisely focusing on misdiagnosed patients, and embracing the complexity in underlying relationships between events and patient covariates. Based on more individualized predictions our hope is that clinicians will improve long term prognosis, even for atypical patients.

2. Background

2.1. Problem Formulation

Our goal is to develop a prognostic risk score for heterogeneous populations. Each patient i is characterized by a d-dimensional vector of covariates $\boldsymbol{x}_i \in \mathcal{X}$, $\mathcal{X}$ a d dimensional input space, an outcome variable $T_i \in \mathbb{R}^+$ which represents the time until occurrence of the event of interest and an indicator variable $\delta_i = I(T_i < C_i)$ that indicates the type of event observed. Patients being followed in a medical study may drop-out resulting in a potential event being unobserved, C_i represents this censoring time. Thus here 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 represents the probability of event occurrence after time t as a function of time t and patient covariates $\boldsymbol{x}_i$,

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

The relationship between patient covariates, time and survival outcome will be complex for many modern data sets. Our aim is to estimate S allowing for *flexible interactions* between patient covariates, time and survival that are *personalized* to each individual. That is, S will be modelled by a flexible function with few assumptions constraining its behaviour. The relationship between survival and patient covariates is to be estimated from an observational data set $\mathcal{D}$ comprising n patients assumed to be drawn *i.i.d.* from the random tuple $\{\boldsymbol{X}_i, \delta_i, \delta_i T_i + (1 - \delta_i)C_i\}$. The probability of event within a suitable time window Δt, $\mathbb{P}(T < t + \Delta t|T > t, \boldsymbol{x})$ is used as a risk score based on which clinicians design therapies for patients.

2.2. Related work

Survival analysis differs from other supervised settings by not only focusing on the event of interest but also analyzing the time to event. A full survival distribution is needed to guide therapy as opposed to single event predictions like those in standard classification settings. Prognostic tools most often build upon the Cox proportional hazards model (Cox, 1972; Katzman et al., 2016) and probabilistic frameworks based on parametric survival distributions (Fernández et al., 2016; Ranganath et al., 2016). The extensions cited above very flexibly model the interactions between a patient's covariates and her survival but issue predictions from parametric functions that restrict the survival behaviour over time. For instance the Weibull distribution used in (Fernández et al., 2016; Ranganath et al., 2016) has a monotonically increasing or decreasing hazard function.

Boosting for survival analysis Boosting based algorithms have been proposed to study survival as extensions to the Cox proportional hazard model in (Ridgeway, 1999; Li and Luan, 2005). These were introduced in the gradient boosting framework (Friedman, 2001) that interprets the process of boosting prediction models as a step-wise optimization procedure applicable to any arbitrary differentiable loss function. The idea is to pursue iterative steepest ascent of the log likelihood function. The work in (Ridgeway, 1999) proposed to update parameter values β, in a linear model, computed based on the negative gradient of Cox's partial likelihood. Subsequently in (Li and Luan, 2005) the authors extended the procedure to handle high-dimensional gene expressions. Alternative parameter (β) optimization procedures have also been proposed in (Binder and Schumacher, 2008) and (Mayr and Schmid, 2014). The former involves the direct maximization of the partial log-likelihood to update parameter values in a linear model, rather than based on correlations as in (Ridgeway, 1999). The latter proposed to optimize a smoothed approximation to the concordance index, a common performance measure for censored outcomes. We note that in all the above mentioned boosting algorithms, final survival estimates are computed with,

$$S(t|\boldsymbol{x}_i) = \exp\left(\int \lambda_0(t)\exp(\hat{\beta}^T\boldsymbol{x}_i)dt\right) \tag{2}$$

where $\lambda_0(t)$ is the baseline hazard function, related to survival by $\lambda(t) = -\partial \log(S(t))/\partial t$, and $\hat{\beta}$ is a vector of estimated parameter values. Therefore they carry the assumption of proportionality of hazards (the ratio of hazards $\lambda_i(t)/\lambda_j(t)$ is independent of time) and assume linearity in covariate interactions ($\hat{\beta}^T\boldsymbol{x}_i$). In contrast our method learns flexible nonparametric survival functions for an individual patient not restricted by proportionality of hazards and is able to learn arbitrary interactions between patient covariates. This is important to capture individual idiosyncrasies and truly provide individualized prognosis (Ahuja et al., 2017).

Bagging for survival analysis Bagging based algorithms include Random Survival Forests methods such as those introduced in (Ishwaran et al., 2008) and (Hothorn et al., 2006). These are *parallel* ensembles in which *fully* grown survival trees are built independently on a bootstrapped sample of the data. In contrast we propose a *sequential* procedure with *shallow* survival trees grown on a data sample dependent on performance of previous trees. Bagging-based algorithms thus do not aim at correcting mistakes of previous hypotheses but aim to decrease overall variability of single tree predictions.

3. Boosted Trees for Risk Prognosis

This section describes our main contribution: two variants of Freund and Shapire's Adaboost algorithm (Freund and Schapire, 1995) for survival prediction we call SurvivalBoost.R and SurvivalBoost.T (R stands for regression and T for threshold). We will first describe the building blocks/steps of our approach in separate sections and then bring those together in a description of the overall procedure.

3.1. Measuring misdiagnoses

The key to boosting architectures is the re-weighting of those patients that are "misclassified" at each iteration. In survival problems, the output given by a hypothesis h for a patient i is not correct or incorrect, but a probability function over time. Labels, when observed, correspond to a draw from an underlying true survival distribution. We propose to measure the "miss-classification" between the model survival predictions and the true survival state of the patient at a given time as the mean squared difference in actual $I(T_i > t)$ and predicted $\hat{h}(t; \boldsymbol{X}_i)$ survival outcomes over time, called the Brier Score (Mogensen et al., 2012). For prediction over the range of all future times we aggregate the Brier Score over time resulting in the Integrated Brier Score (IBS) .

$$IBS(\tau) := \frac{1}{\tau}\int_0^{\tau} \mathbb{E}\left[\left(I(T_i > t) - \hat{h}(t; \boldsymbol{X}_i)\right)^2\right] dt \tag{3}$$

I stands for the indicator function. For censored patients, the time to the event of interest T_i will be unobserved and thus we approximate the integrand by its empirical mean weighted by the inverse probability of censoring at each time t, $\hat{W}_i(t)$ (Mogensen et al., 2012).

3.2. Survival Tree Construction

Trees are composed of leaves and nodes. Leafs define a partition for the data and are responsible for making predictions and nodes guide examples towards appropriate leaves using binary splits based on boolean-valued rules. Each node of our trees h partitions the population in more homogeneous subsets based on the split that results in the greatest reduction in the deviance (a measure of goodness of fit), assuming an exponential likelihood for the data (LeBlanc and Crowley, 1992). For a node C and individuals i in that node, the within-node deviance is defined as:

$$D_C = \sum_{i \in C} \delta_i \log\left(\frac{\delta_i}{\hat{\lambda} t_i}\right) - (\delta_i - \hat{\lambda} t_i) \tag{4}$$

where $\hat{\lambda} := \sum_{i\in C} \delta_i / \sum_{i \in C} t_i$ is the maximum likelihood statistic for the rate parameter in the exponential model. The splitting criterion, as a function of the splitting covariate and cut-off value, then chooses the partition of the population (left and right children nodes in a tree representation) that maximizes the likelihood ratio statistic: $D_{parent} - (D_{left-daughter} + D_{right-daughter})$, which measures the improvement in goodness of fit resulting from this partition. The performance of the reduction in the one-step deviance is very similar to the log-rank test statistic used in other survival tree implementations such as the Random Survival Forest of (Ishwaran et al., 2008) and its performance compares favourably to other splitting methods in the simulation analysis of (Shimokawa et al., 2015) for a variety of underlying hazard behaviours. The deviance has the advantage of quantifying the goodness of fit of a single split with respect to the overall tree which is used to understand the benefit of a single split and therefore also the influence of the covariate used in that split. We discuss covariate influence in section 3.5.

3.3. Terminal node predictions

Terminal node predictions of survival trees are made with the Kaplan-Meier estimator. Let $\mathcal{C}_j$ denote the index set of patients with terminal node j, we compute survival prediction at terminal node j with the Kaplan-Meier estimator,

$$\hat{h}_j(t) = \prod_{i \in \mathcal{C}_j : t_i \leq t} \left(1 - \frac{N_j(t_i)}{Y_j(t_i)}\right) \tag{5}$$

where $N_j(t_i)$ is the number of events at time t_i in terminal node j and $Y_j(t_i)$ is the total number of individuals at risk at time just before t_i in terminal node j. The terminal nodes

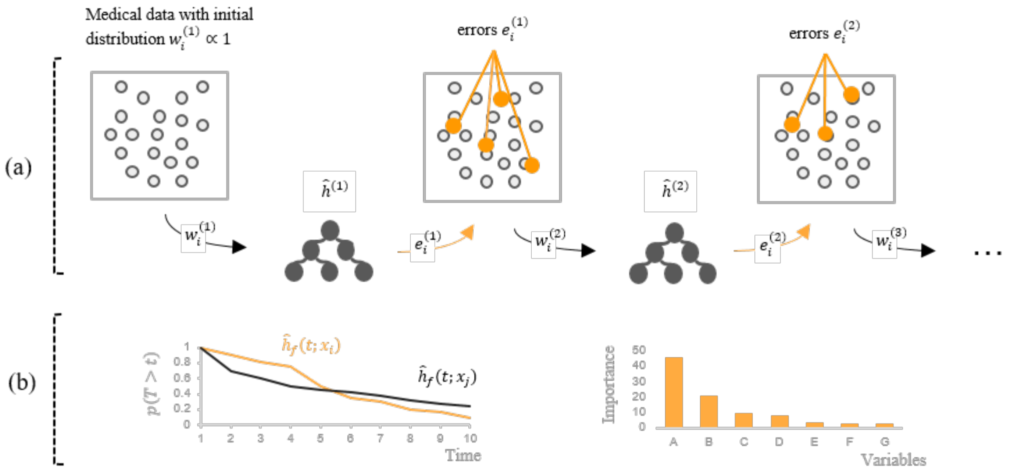

Figure 1: Overview of our boosting procedure for survival analysis. (a) illustrates the training procedure of our model: the second survival tree $\hat{h}^{(2)}$ is trained on weighted samples $w_i^{(2)}$ of the patient data, depending on the predictive performance of previous survival trees on individual patients (in this case through errors $e_i^{(1)}$ of tree $\hat{h}^{(1)}$). (b) shows the final output of the ensemble: we combine predictions (individual survival functions) of learned trees $\hat{h}^{(1)}, \hat{h}^{(2)}, ...$ to produce flexible estimates of survival (left) and investigate influential covariates in the tree construction which results in a measure of covariate importance (right).

partition the sample space so this defines the survival function for the tree,

$$\hat{h}(t; \boldsymbol{x}_i) = \sum_j I(i \in \mathcal{C}_j) \hat{h}_j(t) \tag{6}$$

3.4. Ensemble Model

We introduce two implementations: SurvivalBoost.R and SurvivalBoost.T, which differ in the interpretation of errors or misdiagnoses by individual hypotheses. Our procedure is similar to the regression algorithms Adaboost.R2 by (Drucker, 1997) and Adaboost.RT by (Solomatine and Shrestha, 2004).

- SurvivalBoost.R: The error of hypothesis $\hat{h}$ on each patient e_i is defined in SurvivalBoost.R as the individual estimate of the IBS,

$$e_i := \frac{1}{T} \int_0^T \hat{W}_i(t) \left(I(T_i^* > t) - \hat{h}(t; \boldsymbol{x}_i) \right)^2 dt \tag{7}$$

- SurvivalBoost.T: In contrast SurvivalBoost.T maps individual prediction errors e_i to the set $\{0, 1\}$ by comparing these to a threshold ϕ specified by the user, typically set by cross-validation and thus recovering the familiar classification setting of Adaboost. Specifically for SurvivalBoost.T,

$$e_i := I\left(\frac{1}{T}\int_0^T \hat{W}_i(t)\left(I(T_i^* > t) - \hat{h}(t;\boldsymbol{x}_i)\right)^2 dt > \phi\right) \tag{8}$$

We note that performance results in practice can depend heavily on the choice of the threshold ϕ, as noted in (Shrestha and Solomatine, 2006). We found this drawback to be overcome consistently by heuristically setting ϕ such that initially 30% of the training sample is forced to be considered miss-classified.

In each iteration m of the algorithm, following the error computation e_i for all i in the training set, we assign confidence $\beta^{(m)}$, a function of the average error with respect to the data distribution, to individual hypotheses adjusted to lie in the interval $[0, 1]$ (Line 5 in Algorithm 1). Note that random guessing corresponds to an average error of $e_i = 1/3$. Increased weight is subsequently assigned to patients for which the error is largest, and lowered weight for more accurate predictions; the magnitude of this update is determined by the confidence $\beta^{(m)}$ of an individual tree (line 6 of Algorithm 1). The algorithm is thus encouraged to focus on a potentially different subset of individuals sampled from the training set which become relatively harder to predict. We learn each individual hypothesis on a sub-sample of the original training data drawn with probabilities proportional to the weights updated in the previous round and without replacement. Incorporating randomness as an integral of the procedure decreases computational complexity and tends to reduce correlation in solutions of successive weak learners which we observed empirically to impact performance favourably (see the supplement). A theoretical justification for this observation can be found in the decomposition of the mean squared error with respect to an underlying *true* model $S(t)$. The mean squared error is positively related to the correlation between individual survival trees; see the supplement for a derivation.

Final survival estimates result from a weighted average of individual survival trees $\hat{h}^{(m)}$ with respect to the negative logarithm of their individual confidence, $\log(1/\beta^{(m)})$; low error of individual hypothesis leads to increased weight in the ensemble.

$$\hat{h}_f(t;\boldsymbol{x}_i) := \frac{\sum_{m=1}^{M}\log(1/\beta^{(m)})\hat{h}^{(m)}(t;\boldsymbol{x}_i)}{\sum_{m=1}^{M}\log(1/\beta^{(m)})} \tag{9}$$

A graphical illustration of the boosting procedure is shown in Figure 1. The complete implementation of both algorithms is shown in Algorithm 1.

3.5. Covariate importance

Covariates that are used to make splits that improve goodness of fit are informative of survival predictions in comparison to others with less evidence of improving the fit. The deviance criterion introduced in equation 4 can be used to measure this relevance. Similarly to (Breiman, 2017), we measure overall covariate importance of an individual covariate by examining the sum of the likelihood ratio statistics, measuring goodness of fit, for each split using the covariate of interest in each tree of the ensemble. This is in contrast to Random Survival Forests of (Ishwaran et al., 2008) which uses a covariate permutation approach to identify changes in prediction error due to that covariate.

Algorithm 1 SurvivalBoost

```
Input: Survival data set $\mathcal{D} = \{(X_i, T_i, \delta_i)\}_i$ of size $n$, number of iterations $M$, initial weights $w_i^{(1)} \propto 1$, threshold $\phi$, sampling fraction $s$.
Compute inverse probability of censoring weights $\hat{W}_i(t)$ for each patient $i$.
for $m = 1$ to $M$ do
    1. Let $\mathcal{D}^*$ be a randomly sampled fraction $s$ of training data $\mathcal{D}$ with distribution $w^{(m)}$.
    2. Learn hypothesis $h^{(m)} : \mathcal{X} \times T \to [0, 1]$ on $\mathcal{D}^*$.
    3. Calculate prediction error $e_i^{(m)}$ for each patient $i$.
        - SurvivalBoost.R: with equation 7.
        - SurvivalBoost.T: with equation 8.
    4. Calculate adjusted error of $h^{(m)}$, $\epsilon^{(m)} = \sum_i e_i^{(m)} w_i^{(m)}$.
    5. Calculate confidence in individual hypothesis:
        - SurvivalBoost.R: Let $\beta^{(m)} = \frac{\epsilon^{(m)}}{2/3 - \epsilon^{(m)}}$.
        - SurvivalBoost.T: Let $\beta^{(m)} = \epsilon^{(m)}$.
    6. Update data distribution.
        - SurvivalBoost.R: $w_i^{(m+1)} \propto w_i^{(m)} (\beta^{(m)})^{1 - e_i^{(m)}}$.
        - SurvivalBoost.T: $w_i^{(m+1)} \propto w_i^{(m)} \beta^{(m)} I(e_i^{(m)} = 1) + w_i^{(m)} I(e_i^{(m)} = 0)$.
end for
Output: Final hypothesis $\hat{h}_f$, the weighted average of $\hat{h}^{(m)}$ for $1 \leq m \leq M$ using $\log(1/\beta^{(m)})$ as the weight of hypothesis $\hat{h}^{(m)}$.
```

3.6. Computational Complexity

The computational complexity of both implementations is $\mathcal{O}(DN(M+logN))$, where D is the number of covariates, N the number of patients and M the number of iterations. The burden of the complexity lies in the construction of the survival trees as the rest of the operations can be performed in $\mathcal{O}(N)$. Assuming the data samples are sorted in each covariate the cost of finding the best survival tree is $\mathcal{O}(DN)$. Sorting all the covariates will take $\mathcal{O}(DNlogN)$ time and this has to be done only once before starting the first iteration. Hence, the overall cost of M iterations is $\mathcal{O}(DN(M + logN))$. Now, the re-sampling mechanism can drastically reduce the computational complexity of the proposed approaches as only a fraction of N is used in every iteration.

4. Experimental Setup

Our experiments are presented in 2 parts. We present predictive performance results in comparison to competitive baseline algorithms on 7 different medical data sets related to cardiology. Next we introduce a synthetic data generation scenario to investigate the accurateness of covariate importance summaries and compare the performance of our boosting approach to Random Survival Forest on a challenging subset of the population. In the supplement we provide an analysis of the dependence of performance on the complexity of the ensemble and amount of randomization introduced; these are the two hyper-parameters in our approach.

4.1. Evaluation

Performance assessment. In the presence of censoring we adopt two common approaches used in the literature: the time-dependent concordance index (C-index) (Gerds et al., 2013) defined as,

$$C(t) := \mathbb{P}(\hat{S}_i(t) > \hat{S}_j(t) | \delta_i = 1, t \leq T_j, T_i > T_j) \tag{10}$$

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. We also evaluate performance with the Integrated Brier Score (IBS) introduced in equation 7. In all experiments, these metrics are adjusted for censoring as in (Gerds et al., 2013) and (Mogensen et al., 2012).

Performance comparisons. The most natural algorithmic comparisons are done with existing ensemble methods in the survival analysis literature, those based on bagging and boosting. As a first comparison we evaluate the widely used standard Cox model (Cox) (Cox, 1972) that serves as a semi-parametric baseline with specified covariate interactions but unspecified baseline hazard function. We also compare with the Cox proportional hazards model by component-wise likelihood-based boosting (CBL) from (Binder and Schumacher, 2008) and the model-based boosting algorithm with the implementation based on the work of (Ridgeway, 1999) (CBM); both described in section 2.1. Implementations are done with the R packages `CoxBoost` and `gbm`. We evaluated also the approach of (Mayr and Schmid, 2014) that we denote C-index boosting (CindexBoost), directly maximizing a smoothed version of the C-index. The implementation is done with the R package `mboost` and code provided by the authors. For all boosting algorithms the number of iterations is optimized via cross-validation. Bagging-based algorithms used for comparison are Random Survival Forest (RSF) (Ishwaran et al., 2008) implemented with the R package `RandomForestSRC`, with the number of trees optimized through a grid search; and conditional inference forests (CRSF) (Hothorn et al., 2006) using conditional inference survival trees to construct the ensemble. CRSF is implemented with the R package `pec`.

4.2. Medical data studies

Our experiments on real medical data investigate the discriminative ability of SurvivalBoost.R and SurvivalBoost.T on patients at various stages of the trajectory of cardiovascular diseases (CVD). We consider preventive care efforts for patients at early stage of CVD development, end stage cardiac patients referred for heart transplantation and multimorbid patients diagnosed with cancer but simultaneously at risk of CVD. We give a brief description of the data below and refer the reader to the Supplementary material for more details and summary statistics.

4.2.1. Preventive care

We considered two major cohorts for preventive cardiology. The first is the Meta-analysis Global Group in Chronic heart failure database (MAGGIC), which holds data for $40,366$ patients gathered from multiple clinical studies (Wong et al., 2014). The second cohort was extracted from the UK Biobank, which is a bio-repository with primary care data for more than 500,000 patients in the UK (Sudlow et al., 2015).

4.2.2. Heart transplant wait-list management

We extracted data from the United Network for Organ Sharing (UNOS) database [1], which encompasses an open cohort of prospectively collected data containing information on all patients undergoing heart transplantation in the U.S. We selected a population of 792 patients wait-listed to receive a transplant.

4.2.3. Cancer diagnosed patients

We extracted 2 cohorts from the Surveillance, Epidemiology, and End Results (SEER) cancer registries. SEER is a public database [2] which provides information on cancer diagnosed patients in the U.S. population. We consider patients diagnosed with breast cancer (SEER-I) and leukemia cancer (SEER-II).

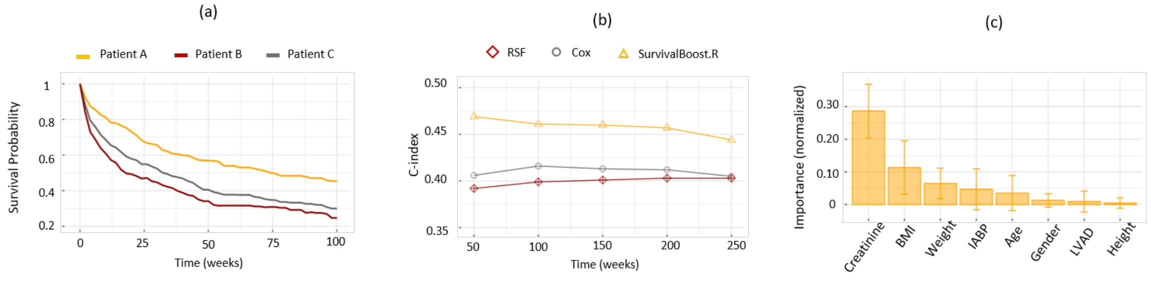

Figure 2: Panel (a) shows an illustration of predicted survival curves with SurvivalBoost.R for three selected patients from the UNOS data set with different levels of Creatinine. In panel (b) we show predictive performance of SurvivalBoost.R, RSF and Cox models on a challenging sub-population of UNOS. In panel (c) wo show predicted covariate importance for the UNOS data according to SurvivalBoost.R.

4.3. Results

Predictive performance is computed by 5-fold cross-validation with implementation as described in section 4.1. Hyper-parameters of SurvivalBoost.R and SurvivalBoost.T were defaulted to 250 trees and a tree depth of 3 which were found to perform consistently well.

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

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

Models	**UNOS**	**MAGGIC**	**UK Bio.**	**SEER-I**	**SEER-II**
Cox	0.603 ± 0.04	0.645 ± 0.01	0.679 ± 0.02	0.772 ± 0.03	0.740 ± 0.03
CBL	0.605 ± 0.04	0.644 ± 0.01	0.679 ± 0.02	0.774 ± 0.03	0.738 ± 0.04
CBM	0.635 ± 0.03	0.625 ± 0.01	0.673 ± 0.02	0.768 ± 0.03	0.740 ± 0.04
CindexBoost	0.564 ± 0.06	0.592 ± 0.01	0.655 ± 0.03	0.764 ± 0.03	0.742 ± 0.04
SRF	0.634 ± 0.04	0.642 ± 0.01	0.627 ± 0.01	0.686 ± 0.03	0.680 ± 0.01
CSRF	0.635 ± 0.05	0.652 ± 0.02	0.638 ± 0.02	0.755 ± 0.03	0.717 ± 0.04
SurvivalBoost.R	0.636 ± 0.03	0.676 ± 0.02	0.702 ± 0.02	**0.780 ± 0.03**	**0.752 ± 0.03**
SurvivalBoost.T	**0.647 ± 0.04**	**0.675 ± 0.04**	**0.725 ± 0.03**	0.775 ± 0.04	0.740 ± 0.04

Table 1: C-index figures (higher better) and standard deviations on all data sets.

Models	**UNOS**	**MAGGIC**	**UK Bio.**	**SEER-I**	**SEER-II**
Cox	0.204 ± 0.02	0.177 ± 0.01	**0.013 ± 0.00**	**0.042 ± 0.01**	0.054 ± 0.00
CBL	0.202 ± 0.02	0.177 ± 0.01	**0.013 ± 0.00**	0.043 ± 0.01	0.054 ± 0.00
CBM	0.190 ± 0.01	0.179 ± 0.01	0.016 ± 0.00	0.044 ± 0.01	0.054 ± 0.00
CindexBoost	0.210 ± 0.02	0.181 ± 0.01	0.016 ± 0.00	0.044 ± 0.01	**0.053 ± 0.00**
SRF	0.199 ± 0.02	0.176 ± 0.02	0.014 ± 0.00	0.050 ± 0.01	0.059 ± 0.00
CSRF	0.193 ± 0.01	0.175 ± 0.01	0.014 ± 0.00	0.045 ± 0.01	0.057 ± 0.00
SurvivalBoost.R	0.186 ± 0.01	0.162 ± 0.01	**0.013 ± 0.00**	**0.042 ± 0.00**	0.054 ± 0.00
SurvivalBoost.T	**0.185 ± 0.01**	**0.160 ± 0.01**	**0.013 ± 0.00**	0.044 ± 0.00	0.055 ± 0.00

Table 2: Integrated Brier Score (lower better) and standard deviations on all data sets.

For SurvivalBoost.R and SurvivalBoost.T a sub-sampling fraction of 80% and 50% was used, suggested by the synthetic analysis provided in the Supplement.

We illustrate inference based on our model with the UNOS data set on Figure 2. On panel (a) we show a sample of predicted survival trajectories for three selected patients with different levels of creatinine; a biomarker found to be highly relevant for disease progression with our covariate importance procedure (panel (c)). Creatinine has been shown to adversely impact progression of cardiovascular disease in (Wannamethee et al., 1997) which is reflected in our findings; patient A, B and C have measured creatinine levels in the $95^{th}, 50^{th}$ and 5^{th} percentile of the population respectively. On panel (b) we show C-index results at different time horizons on a subset of challenging patients of UNOS which we defined as: the 1/3 of the UNOS population with highest Integrated Brier Score (the subset with the highest discrepancy between predicted and actual survival) for survival predicted with a simple survival tree. In comparison to Cox and RSF, SurvivalBoost.R significantly outperforms on this subset which suggests that based on our model we would able to improve outcomes on patients that would otherwise be consistently mistreated. We get similar results with SurvivalBoost.T.

Tables 1 and 2 show performance measured by the C-index at the 50% percentile of the empirical event time distribution and the Integrated Brier Score for all algorithms on all experiments. SurvivalBoost.R and SurvivalBoost.T are superior on all data sets with respect to the bagging-based algorithms SRF and CSRF which suggests that overall prognosis can be efficiently improved by boosting in contrast to a parallel bagging approach. We note also large performance gains with respect to the Cox-based models for the UNOS, UK Biobank and MAGGIC data sets for which non-linear interactions influence survival; which can be seen also by the competitive performance of the SRF and CSRF nonparametric algorithms. Our results show that SurvivalBoost.T tends to outperform SurvivalBoost.R on these data sets with patients at risk of cardiovascular diseases for which disease progression dynamics may be more heterogeneous and complex than for cancer-related SEER. We believe that on these data sets, the coarse threshold ϕ used by SurvivalBoost.T acts as an implicit regularizer, more robust to outliers which may occur in heterogeneous cohorts. On the SEER data sets Cox provides highly competitive performance which suggests that a linear combination of covariates provides a good description of survival. In this case the more subtle relative weight update measure (equation 8) of SurvivalBoost.R appears to better capture the more subtle differences between patients in comparison to SurvivalBoost.T.

4.4. Simulation Studies

We illustrate the ability of SurvivalBoost to distinguish between the relevance of different covariates with a synthetically generated population. We consider a non-linear data generating process based on the following association rule,

$$\Lambda(\boldsymbol{x}_i) := 4 + \log(0.1x_{i,1} + 0.2x_{i,2} + 0.3x_{i,3}) + x_{i,4}$$

To ensure dynamics approximating survival settings $\Lambda(\boldsymbol{x})$ determines the shape parameter in a Weibull distribution ($\mathcal{W}$) as in Table 3.

Covariates	**Time to event**	**Censoring**
$\boldsymbol{X}_i \sim \mathcal{U}(0,1)$	$T_i \sim \mathcal{W}(2, \exp(\Lambda(\boldsymbol{X}_i)))$	with prob. 0.8, $C_i \leftarrow \mathcal{U}(0, T_i)$

Table 3: Synthetic Data Generation

Each individual is described by 5 real valued covariates independently sampled from a standard uniform distribution but only 4 of them influence survival. We combine irrelevant and informative covariates to mimic information recorded in real world medical settings. In addition, we introduce imbalance in the outcome distribution by inducing right-censoring on a random subset of approximately 80% of individuals by altering survival time as follows: $C_i \leftarrow \mathcal{U}(0, T_i)$. This is to reproduce a setting in which the event of interest is rare. We generated 10 data sets of 500 patients by sampling covariates and coefficients with these settings.

Figure 3 shows that SurvivalBoost in both implementations is able to recover the relative influence of covariates successfully. Note the increasing influence on survival of covariates X_1,X_2 and X_3 (since the coefficients are set to 0.1, 0.2 and 0.3) which is reflected in our results. X_5 was introduced as noise with no influence on survival.

5. Conclusions

We have introduced two boosting-based algorithms for survival prognosis and inference for covariate importance, designed to handle the heterogeneity present in modern medical datasets. Traditional survival analysis poses restricting assumptions on the data-generating process, forcing latent patterns to conform to prior assumptions regarding patient behaviour. Our approach overcomes this challenge through the use of an agnostic nonparametric framework. Our experiments on synthetic data suggest that both algorithms are able to correctly infer the relevance of covariates in a population of nonlinear survival dynamics. With extensive evaluations on real medical data, we have demonstrated performance improvements over current techniques.

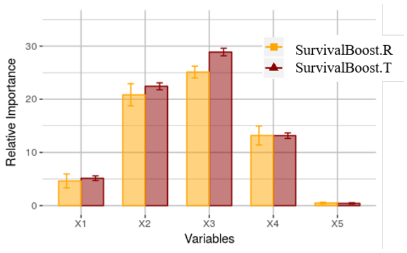

Figure 3: Predicted covariate importance for SurvivalBoost.R and SurvivalBoost.T.

References

Kartik Ahuja, William Zame, and Mihaela van der Schaar. Dpscreen: Dynamic personalized screening. In *Advances in Neural Information Processing Systems*, pages 1321–1332, 2017.

Emelia J Benjamin, Michael J Blaha, Stephanie E Chiuve, Mary Cushman, Sandeep R Das, Rajat Deo, J Floyd, M Fornage, C Gillespie, CR Isasi, et al. Heart disease and stroke statistics-2017 update: a report from the american heart association. *Circulation*, 135(10): e146–e603, 2017.

Harald Binder and Martin Schumacher. Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models. *BMC bioinformatics*, 9(1):14, 2008.

Leo Breiman. *Classification and regression trees*. Routledge, 2017.

David R. Cox. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society. Series B*, 34:187–220, 1972.

Harris Drucker. Improving regressors using boosting techniques. In *ICML*, volume 97, pages 107–115, 1997.

Kenneth H Falchuk and Evan Falchuk. The misdiagnosis epidemic: Five root causes and the growing demand for more patient-centric care. *International Journal of Healthcare Management*, 5(1):61–65, 2012.

Tamara Fernández, Nicolás Rivera, and Yee Whye Teh. Gaussian processes for survival analysis. In *Advances in Neural Information Processing Systems*, pages 5021–5029, 2016.

Yoav Freund and Robert E Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In *European conference on computational learning theory*, pages 23–37. Springer, 1995.

Yoav Freund, Robert Schapire, and Naoki Abe. A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 14(771-780):1612, 1999.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

Thomas A Gerds, Michael W Kattan, Martin Schumacher, and Changhong Yu. Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. *Statistics in Medicine*, 32(13):2173–2184, 2013.

Torsten Hothorn, Kurt Hornik, and Achim Zeileis. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*, 15 (3):651–674, 2006.

Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, and Michael S Lauer. Random survival forests. *The annals of applied statistics*, pages 841–860, 2008.

Sekar Kathiresan and Deepak Srivastava. Genetics of human cardiovascular disease. *Cell*, 148(6):1242–1257, 2012.

Jared Katzman, Uri Shaham, Jonathan Bates, Alexander Cloninger, Tingting Jiang, and Yuval Kluger. Deep survival: A deep cox proportional hazards network. *arXiv preprint arXiv:1606.00931*, 2016.

Michael LeBlanc and John Crowley. Relative risk trees for censored survival data. *Biometrics*, pages 411–425, 1992.

Hongzhe Li and Yihui Luan. Boosting proportional hazards models using smoothing splines, with applications to high-dimensional microarray data. *Bioinformatics*, 21(10):2403–2409, 2005.

W Robb MacLellan, Yibin Wang, and Aldons J Lusis. Systems-based approaches to cardiovascular disease. *Nature Reviews Cardiology*, 9(3):172, 2012.

Andreas Mayr and Matthias Schmid. Boosting the concordance index for survival data–a unified framework to derive and evaluate biomarker combinations. *PloS one*, 9(1):e84483, 2014.

Ulla B Mogensen, Hemant Ishwaran, and Thomas A Gerds. Evaluating random forests for survival analysis using prediction error curves. *Journal of statistical software*, 50(11):1, 2012.

Gene R Quinn, Darrell Ranum, Ellen Song, Margarita Linets, Carol Keohane, Heather Riah, and Penny Greenberg. Missed diagnosis of cardiovascular disease in outpatient general medicine: Insights from malpractice claims data. *Joint Commission journal on quality and patient safety*, 43(10):508–516, 2017.

Rajesh Ranganath, Adler Perotte, Noémie Elhadad, and David Blei. Deep survival analysis. In *Proceedings of the 1st Machine Learning for Healthcare Conference*, volume 56, pages 101–114. PMLR, 18–19 Aug 2016.

Greg Ridgeway. The state of boosting. *Computing Science and Statistics*, pages 172–181, 1999.

Renate B Schnabel, Lisa M Sullivan, Daniel Levy, Michael J Pencina, Joseph M Massaro, Ralph B D'Agostino Sr, Christopher Newton-Cheh, Jennifer F Yamamoto, Jared W Magnani, Thomas M Tadros, et al. Development of a risk score for atrial fibrillation (framingham heart study): a community-based cohort study. *The Lancet*, 373(9665): 739–745, 2009.

Asanao Shimokawa, Yohei Kawasaki, and Etsuo Miyaoka. Comparison of splitting methods on survival tree. *The international journal of biostatistics*, 11(1):175–188, 2015.

Durga L Shrestha and Dimitri P Solomatine. Experiments with adaboost.rt, an improved boosting scheme for regression. *Neural computation*, 18(7):1678–1710, 2006.

Thomas R Skinner, Ian A Scott, and Jennifer H Martin. Diagnostic errors in older patients: a systematic review of incidence and potential causes in seven prevalent diseases. *International journal of general medicine*, 9:137, 2016.

Dimitri P Solomatine and Durga L Shrestha. Adaboost.rt: a boosting algorithm for regression problems. In *Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on*, volume 2, pages 1163–1168. IEEE, 2004.

Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3):e1001779, 2015.

S Goya Wannamethee, A Gerald Shaper, and Ivan J Perry. Serum creatinine concentration and risk of cardiovascular disease. *Stroke*, 28(3):557–563, 1997.

Chih M Wong, Nathaniel M Hawkins, Mark C Petrie, Pardeep S Jhund, Roy S Gardner, Cono A Ariti, Katrina K Poppe, Nikki Earle, Gillian A Whalley, Iain B Squire, et al. Heart failure in younger patients: the meta-analysis global group in chronic heart failure (maggic). *European heart journal*, 35(39):2714–2721, 2014.

Chen Yao, Brian H Chen, Roby Joehanes, Burcak Otlu, Xiaoling Zhang, Chunyu Liu, Tianxiao Huan, Oznur Tastan, L Adrienne Cupples, James B Meigs, et al. Integromic analysis of genetic variation and gene expression identifies networks for cardiovascular disease phenotypes. *Circulation*, 2014.