SURVFIT: Doubly Sparse Rule Learning for Survival Data

Ameer Hamza Shakur
Industrial and Systems Engineering, University of Washington - Seattle, ahamza@uw.edu

Shuai Huang
Industrial and Systems Engineering, University of Washington - Seattle, shuaih@uw.edu

Xiaoning Qian
Electrical and Computer Engineering, Texas A&M University, xqian@tamu.edu

Xiangyu Chang
School of Management, Xi'an Jiaotong University, xiangyuchang@xjtu.edu.cn

Survival data analysis has been leveraged in medical research to study disease morbidity and mortality, and to discover significant bio-markers affecting them. A crucial objective in studying high dimensional medical data is the development of inherently interpretable models that can efficiently capture sparse underlying signals while retaining a high predictive accuracy. Recently developed rule ensemble models have been shown to effectively accomplish this objective; however, they are computationally expensive when applied to survival data and do not account for sparsity in the number of variables included in the generated rules. To address these gaps, we present SURVFIT, a "doubly sparse" rule extraction formulation for survival data. This doubly sparse method can induce sparsity both in the number of rules and in the number of variables involved in the rules. Our method has the computational efficiency needed to realistically solve the problem of rule-extraction from survival data if we consider both rule sparsity and variable sparsity, by adopting a quadratic loss function with an overlapping group regularization. Further, a systematic rule evaluation framework that includes statistical testing, decomposition analysis and sensitivity analysis is provided. We demonstrate the utility of SURVFIT via experiments carried out on a synthetic dataset and a sepsis survival dataset from MIMIC-III.

Key words: rule learning, survival analysis, sparsity, second-order cone programming

1. Introduction

When analyzing biological and medical datasets, an often encountered scenario is the need to simultaneously analyze multiple variables and understand their impact on a certain disease or biological condition. In this endeavor, regression methods have been a typical approach. These methods help us understand the relative importance of variables primarily in terms of their average effects on the outcome rather than their synergistic interactions. Though adding interaction terms to the regression model can certainly enable their application in evaluating the significance of these interaction terms, regression models themselves

are not adequate for discovering such interactions due to both computational and statistical challenges, i.e., the number of potential interactions grows at a super-exponential rate regarding the number of variables. The rule learning approach is a natural way to address these challenges. An old song since its inception in the early 70s and 80s as a typical approach of Artificial Intelligence, it now finds its new tune in the 21st century as a result of considerable developments in the fields of machine learning and optimization such as random forests (Breiman 2001) and sparse regularization models such as LASSO (Tibshirani 1996). Rulefit (Friedman and Popescu 2008) is a good example of a model that cleverly combines these methods by first generating a huge list of rules from a tree ensemble, and then applying LASSO to select a minimum set of rules that can predict the outcome with a good accuracy. Compared to rule learning methods developed before Rulefit that mostly used heuristic algorithms (Cendrowska 1987, Cohen 1995) or logic deduction approaches (Michalski 1980, Quinlan 1990) to derive rules, Rulefit is both computationally efficient, inherited from random forest and LASSO, and statistically well justified, as random forest uses bootstrap aggregation to generate an ensemble of tree models and has the ability to cover a wider range of the rule space, therefore being less susceptible to being stuck in local optima. An additional advantage of applying rule based models to biomedicine is that they can be easily communicated to, and evaluated by medical professionals. Several recent works have successfully applied rule based models to diverse biomedical datasets to understand risk-predictive profile patterns and build predictive models for diseases, including Type 1 diabetes (Lin et al. 2014), Type 2 diabetes (Patil et al. 2010), depression (Lin et al. 2018), classification of cancer gene expression data (Glaab et al. 2012) etc. However, these works were not focused on survival data. Survival rule models proposed in literature (Fokkema 2017, Wróbel et al. 2017) lack the methodology to impose sparsity on the variables that constitute the rules. Sparsity in variables has been proven to be a main concern in a wide range of applications. Surprisingly, sparsity of variables involved in the rules has not yet been addressed in rule learning literature. Therefore, our research seeks to address these gaps and focuses on a rule learning approach that can efficiently learn a "doubly sparse" set of rules and analyze their properties for survival analysis, a field with critical applications in biomedicine.

1.1. Background

Survival analysis is a classical field of statistical learning that has been widely used to study how statistical factors influence morbidity and mortality for different diseases, e.g. congestive heart failure (Paulon et al. 2020), gene selection and screening for lymphoma (Pang et al. 2012) and pediatric trauma (Mittal et al. 2014), to name a few. Challenges in survival analysis applied to medical data stem from complexity of underlying processes, high dimensionality of datasets, and the incompleteness of time to event data.

1.1.1. Survival Analysis. Survival data is indexed by the sequence $(t_i, \delta_i, \mathbf{x}_i)$ for $i \in \{1, \ldots, N\}$ where N is the total number of observations and t_i is the observation time, i.e., $t_i = \min(T_i, C_i)$, where T_i is the time-to-event occurrence (event time) and C_i is the time of last observation (censoring time). The binary variable δ_i represents the status of the i-th observation at the observation time, i.e., δ_i takes value 1 if the event has occurred at t_i, otherwise it takes value 0. Since event times are only available for a small subset of the total observations, the study of survival data is a challenging task. Survival analysis methods are used to model the *survival function*, $S(t; \mathbf{x})$ denoting the probability that the event has not yet occurred at time t for an observation with variables $\mathbf{x}$. Assuming that the time-to-event, T, is a continuous random variable with a probability density function $f(t; \mathbf{x})$, we can define the survival function, $S(t; \mathbf{x})$, as

$$S(t; \mathbf{x}) = \Pr\{T \geq t\} = \int_t^\infty f(s; \mathbf{x}) ds. \tag{1}$$

The *hazard function*, $h(t; \mathbf{x})$ is the instantaneous rate of occurrence of the event at time t that is then defined as

$$h(t; \mathbf{x}) = \lim_{dt \to 0} \frac{\Pr\{t \leq T < t + dt \mid T \geq t\}}{dt} = \lim_{dt \to 0} \frac{\Pr\{t \leq T < t + dt, T \geq t\}}{\Pr\{T \geq t\} dt} = \frac{f(t; \mathbf{x})}{S(t; \mathbf{x})}. \tag{2}$$

Survival analysis is a mature field which includes several standard parametric, semi-parametric, and non-parametric methods as well as modern machine learning models. In the popular Cox regression model (Cox 1972), a proportional hazards assumption that the effects of the predictor variables upon survival are constant over time is made. Cox regression, like other parametric generalized linear models, assumes a specific linear link between the predictor variables and the hazard function such that the ratio of hazards between two observations remains constant over time. Interactions between variables may

be incorporated in this model, but they need to be done explicitly. Thus, machine learning models such as survival trees (LeBlanc and Crowley 1992) and tree ensembles, such as random survival forests (Ishwaran et al. 2008), have been developed to mitigate the limitations of parametric models. Survival trees are a flexible approach to deal with these challenges as they make no assumptions on the response function and can detect interactions automatically.

1.1.2. Survival Trees. Parametric (and semi-parametric) regression models impose a specific link function on the response and face challenges in incorporating interactions between variables. Trees provide a flexible approach that can detect interactions in variables without explicitly specifying them beforehand. They also do not assume a specific link function and are widely used as they are easy to interpret and understand for medical professionals. Trees naturally group together observations with similar outcomes which leads them to be highly interpretable. The main difference between classical decision trees and survival trees is in the splitting criteria used to partition the data - survival trees use criteria that make each child node to be most similar in terms of their survival or hazard functions. Several splitting criteria have been developed for survival trees, e.g., the maximum log-rank statistic (Segal 1988, LeBlanc and Crowley 1992) and the log-rank score (Hothorn and Lausen 2003).

1.1.3. Random Survival Forests. Random forests are an ensemble model generated by combining many decision or survival trees where each tree is built on a randomly bootstrapped sample of data and a randomly selected subset of variables. The average outcome of all of these binary trees is the output of the random forest. Random forests were initially developed for regression and classification problems (Breiman 2001) and later extended to apply to survival data, (Ishwaran et al. 2008, Wright and Ziegler 2017) where an ensemble of survival trees is used to build a forest. For a given input, an average of the cumulative hazard prediction of all the survival trees in the ensemble is the output of the random survival forest.

1.1.4. Sparse Regularization. Modern medical datasets are high dimensional, which leads to challenges associated with the *curse of dimensionality*, particularly in datasets with many correlated variables. This makes it critical to build models that are sparse with respect to the number of variables, and to identify the most significant variables affecting

the underlying process. Development of sparse regularization methods for survival analysis is a line of efforts seeking to deal with the challenges of high dimensional data in both regression-based and tree-based methodologies. Given a data matrix $\mathbf{X} \in \mathbb{R}^{n \times m}$ of n observations and m variables where m is large, our goal in variable selection is to choose a number of variables $k(\ll m)$ that are the most significant predictors of the output. Consider a model parametrized by the coefficients corresponding to the variables, $\boldsymbol{\beta}$ and a loss function $\mathcal{L}(\boldsymbol{\beta}; \mathbf{X})$ that is to be minimized to obtain model coefficient estimates. Sparse regularization works by regularizing the loss function with sparsity inducing norms such as the ℓ_1 norm, $\|\boldsymbol{\beta}\|_1 = \sum_{i=1}^{n} |\beta_i|$ that was first proposed as the LASSO model (Tibshirani 1996) for linear regression. The ℓ_1 norm has the property of shrinking the coefficients closer to zero which enables variable selection by eliminating those variables whose coefficients are nearly zero. An important direction in sparse regularization is structured sparsity regularization to obtain desired model characteristics such as selection of groups of variables, i.e., selecting all variables in a predefined group of variables or none at all. The Group LASSO (Yuan and Lin 2006) solves the group selection problem by using an $\ell_{2,1}$ norm regularization, $\sum_{g=1}^{|G|} \|\boldsymbol{\beta}_g\|_2$, where $\boldsymbol{\beta}_g$ are coefficients in group g, belonging to a set of groups G and the ℓ_2 norm is given by $\|\boldsymbol{\beta}_g\|_2 = \sqrt{\sum_{i \in g} \beta_i^2}$. The Sparse Group LASSO (Simon et al. 2013a) generalizes the Group LASSO to also induce within group sparsity in the solution by using a regularization that is the sum of $\ell_{2,1}$ and ℓ_1 norm, i.e., $\sum_{g=1}^{|G|} \|\boldsymbol{\beta}_g\|_2 + \|\boldsymbol{\beta}\|_1$. As the complexity of the sparsity-inducing norms increases, their adoption to survival regression models such as the Cox regression (Cox 1972) still poses significant algorithmic challenges despite the computational advantages of these methods in the typical regression setting. While most of these efforts mainly focus on variable selection when the link function of the model is linear, the advances in sparse regularization approaches also positively impacted the work on tree-based methods in regression and classification like Rulefit (Friedman and Popescu 2008) as well as in survival analysis, such as the method developed in **pre** (Fokkema 2017), where a sparse set of survival rules are generated by constructing an ℓ_1-regularized Cox regression model over an exhaustive set of rules extracted from the data through bootstrapped survival trees. However, the Cox partial likelihood function used in **pre** has difficulty in scaling to high dimensional data. Although the regularized Cox model can handle relatively high dimensions, the optimization algorithms that are built on the Cox partial likelihood function scale poorly when regularized with structured norms such as Group LASSO and Sparse Group LASSO (Simon et al. 2013b).

1.2. Our Contributions

We propose a new rule learning method, SURVFIT, with three main contributions. First, we aim to fill in a gap that concerns rule learning with variable sparsity, i.e., "double sparsity" for survival data analysis. To achieve this, we propose a formulation that adopts a quadratic loss function and an overlapping group regularization term. The quadratic loss function allows us to bypass the partial log-likelihood loss function of the Cox models that has caused considerable computational difficulty for high-dimensional applications, and the proposed regularization enables us to not just select the most important rules but also induce sparsity of variables involved in the selected rules. This "double sparsity", in both rules and variables, has so far not been addressed in the literature of rule learning. Second, we propose and compare different optimization strategies for solving our optimization problem and discuss their advantages and trade-offs. Third, we provide a systematic rule evaluation framework for evaluating and examining the statistical significance of the rules extracted via SURVFIT. The framework includes statistical testing of rules' ability to discriminate between low risk and high risk observations, decomposition analysis, and sensitivity analysis of the cutoff values. An overall presentation of this framework is shown in Fig. 1. The rest of this paper is organized as follows: Section 2 will introduce the details of SURVFIT, derive the optimization strategy and algorithms, and Section 3 will describe the rule analysis framework. In Section 4, we will present both simulation studies to examine sparsity properties of our method, and a comprehensive data analysis of a medical dataset using SURVFIT. Section 6 will summarize our contributions and conclude this paper. Note that, in this paper, we use lower- or upper-case letters, e.g., x or X, to represent scalars, bold-face lower-case letters, e.g., $\mathbf{x}$, to represent vectors, bold-face upper-case letters, e.g., $\mathbf{X}$, to represent matrices, and upper-case italic letters, e.g., X, to represent random variables.

2. Methodology

Rule learning is a challenging problem mainly due to the combinatorial nature of rules, i.e., a rule is expressed as the product of a few indicator functions $I(\cdot)$ on propositions of values taken by variables in an observation $\mathbf{x}$,

$$r_k(\mathbf{x}) = \prod_{p=1}^{|\mathbf{x}|} I(x_p \in s_{pk}). \tag{3}$$

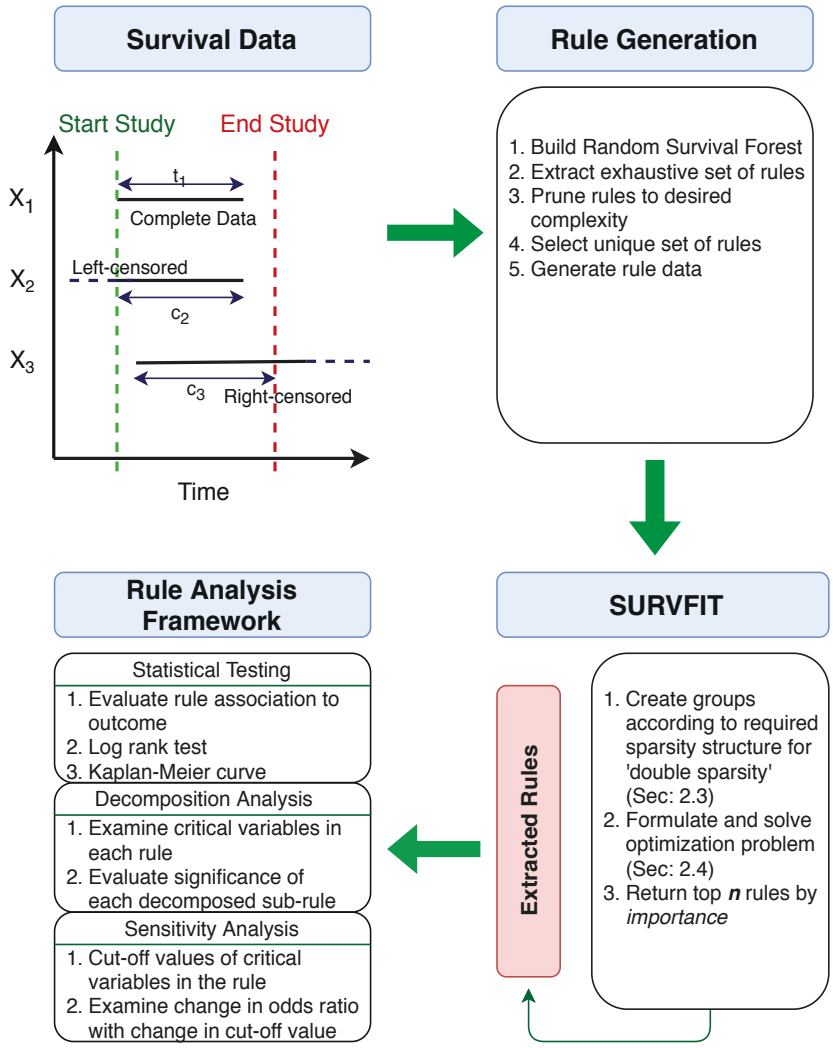
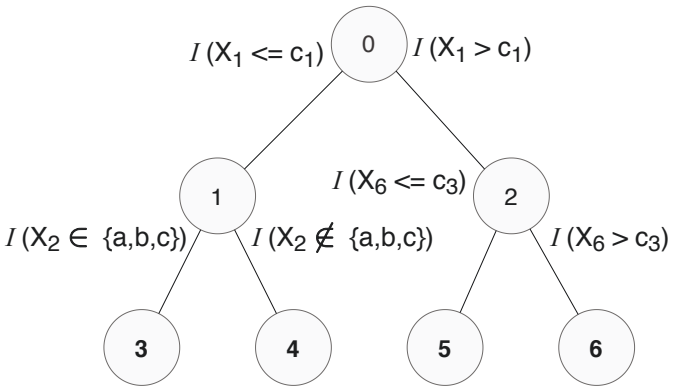


Figure 1 **A schematic outline of the SURVFIT algorithm**

For continuous variables, s_{pk} is a contiguous interval while for categorical variables, it is an explicitly specified set. A rule either gives 0 or 1 as its outcome for an input observation. If its outcome is 1, it means all the conditions on its constituent variables are satisfied, i.e., $\{x_p \in s_{pk}\}_1^{|\mathbf{x}|}$. We say that rule r is endorsed by observation $\mathbf{x}$ if $r(\mathbf{x}) = 1$. Through this combinatorial nature, rules provide an effective semantics to capture interactions among variables, not only in the qualitative sense, i.e., which variables interact with which, but also in the quantitative sense, i.e., the cutoff values used in the conditions of the rules. It is also due to this combinatorial nature that rules are information-rich, but computationally and statistically challenging to detect from data. Recent breakthroughs in rule learning benefit from an insight that a decision tree can be readily decomposed into a set of rules as shown in Fig. 2. Tree ensemble models such as random forests can therefore be used to

generate a huge set of rules. Then, formulations could be developed to filter this set and select a sparse set of the most representative and informative rules. Rule learning methods such as **Rulefit** (Friedman and Popescu 2008) and **pre** (Fokkema 2017) follow this line. However, these methods do not consider the sparsity in the variables that are involved in the extracted rules. Sparsity of variables have proven to be a critical trait of machine learning models that can achieve robust prediction performance and interpretability in practice. An immediate example that will be shown in the medical application in this paper is that variables collected in healthcare applications are usually highly correlated, and thus it is important to be able to generate rules that involve only a sparse selection of significant variables. For example, two variables may show up in different rules, though only one variable is truly significant, and the other is redundant.



NODE	EXTRACTED RULES
3	$X_1 <= c_1$ & $X_2 \in \{a,b,c\}$
4	$X_1 <= c_1$ & $X_2 \notin \{a,b,c\}$
5	$X_1 > c_1$ & $X_6 <= c_3$
6	$X_1 > c_1$ & $X_6 > c_3$

Figure 2 **An example decision tree and corresponding rules extracted from terminal nodes**

2.1. Rule generation

In order to generate an abundant set of rules that will be pruned by our learning formulation, we use the following algorithm to build the random survival forests.

1. Draw a given number of bootstrap samples from the original data.
2. Grow a survival tree for each bootstrapped sample as follows:

(a) Use one of the splitting criteria discussed in 1.1.2 to recursively build a tree using a randomly selected subset of variables for each split.

(b) Grow the tree until no new child nodes can be formed because of the stopping condition that each node must contain a minimum number of unique events.

3. Aggregate all the survival trees to obtain an ensemble.
4. Extract rules of the desired length and complexity from the tree ensemble to generate a large rule list.

2.2. The loss function

This initial set of rules is denoted as $\{r_k(\mathbf{x})\}_1^K$, where K is the total number of rules. We then develop a learning formulation to guide the selection of the final rules, which should be a minimum number of rules (i.e., the number should be much smaller than K) that could achieve optimal prediction on the time-to-event outcome of survival data. It is tempting to use the existing Cox proportional hazards regression model and take the K rules as K input variables, then conduct sparse learning on this Cox model-based formulation. This is a reasonable approach, but the partial likelihood function used in the Cox regression model has been found to scale poorly in high-dimensional applications, particularly with complex group norms (Simon et al. 2013b), including the recently developed **pre** (Fokkema 2017), a rule learning method for survival analysis, that is also built on the negative partial log-likelihood loss in the Cox proportional hazards model. Thus, we resort to another loss function. In this paper, we concern the linear model, but our method could be extended to nonlinear models as well. The prediction by the linear model is,

$$t(\mathbf{x}_i) = \boldsymbol{\beta}_0 + \sum_{k=1}^{K} \boldsymbol{\beta}_k r_k(\mathbf{x}_i). \tag{4}$$

We adopt the robust loss function developed in Ke et al. (2017),

$$\min_{\boldsymbol{\beta}} l(\boldsymbol{\beta}, \mathbf{X}) := \sum_{\{i|\delta_i=1\}} \frac{1}{2}\big(t(\mathbf{x}_i) - t_i\big)^2 + \sum_{\{i|\delta_i=0\}} \frac{\gamma}{2}\big(\min(0, t(\mathbf{x}_i) - t_i)\big)^2. \tag{5}$$

The first term in (5) is the least-squared loss that penalizes the difference between the predicted outcome $t(\mathbf{x}_i)$ and the real outcome t_i for each complete observation $\mathbf{x}_i$. The second term is a squared hinge loss which penalizes the predictions for censored data only when the predicted event time $t(\mathbf{x}_j)$ of censored observation $\mathbf{x}_j$ is smaller than the censor time t_j. The penalty is zero when $t(\mathbf{x}_j)$ is greater than t_j. The hyperparameter γ controls the influence of censored data in parameter estimation and is selected via cross-validation.

2.3. Doubly Sparse Rule Extraction

Consider the regression optimization problem (5) applied to rules, i.e, let $\boldsymbol{\beta}$ be the coefficients of the complete set of rules, and $\mathbf{X}$ the rule data matrix. Each column in $\mathbf{X}$ is a binary variable denoting whether or not the observations endorse a rule. To achieve sparsity in both rules and variables, we integrate two regularization terms simultaneously. On one hand, to enforce sparsity on the rules, we adopt the ℓ_1 norm that was used in the least

absolute shrinkage and selection operator (LASSO) regularization for regression proposed by Tibshirani (1996), i.e., it regularizes the loss function with ℓ_1 norm penalty on the coefficients corresponding to rules ($\|\boldsymbol{\beta}\|_1 = \sum_{i=1}^{n} |\boldsymbol{\beta}_i|$) to push some coefficients to 0. In the rule learning literature for regression, **Rulefit** (Friedman and Popescu 2008) and **pre** (Fokkema 2017) utilise LASSO to extract a sparse rule list from a rule ensemble. Following this line, we propose the following formulation to enforce sparsity in the *cardinality* of rules

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\ l(\boldsymbol{\beta}, \mathbf{X}) + \lambda \cdot \|\boldsymbol{\beta}\|_1, \tag{6}$$

where $l(\boldsymbol{\beta}, \mathbf{X})$ is the squared error loss for regression. As we have mentioned, this formulation does not enforce sparsity among the variables involved in the extracted rules. As one of the main contributions of this work, that is to enforce a "doubly sparse" set of rules, our idea is inspired from the work on sparsity-inducing norms for structured variable selection (Jenatton et al. 2011). A particularly useful concept is the Overlapping Group LASSO regularization. Here, as Fig. 3 shows, the rules form a group structure due to their overlapping use of the variables. To exploit this structural property in order to enforce sparsity on the variables, the regularization term we propose to use is therefore,

$$\Omega(\boldsymbol{\beta}) = \sum_{G \in \mathcal{G}} \|w^G \circ \boldsymbol{\beta}\|_2, \tag{7}$$

where $\mathcal{G}$ is a set of overlapping groups of coefficients; $w^G_{G \in \mathcal{G}}$ are $|\boldsymbol{\beta}|$-dimensional vectors such that $w^G_j > 0$ if $j \in G$ and $w^G_j = 0$ otherwise; $\mathbf{x} \circ \mathbf{y}$ denotes element-wise multiplication of two vectors, $\mathbf{x}$ and $\mathbf{y}$. This regularization term $\Omega(\boldsymbol{\beta})$, despite its difference from the ℓ_1 norm from the surface, is like an ℓ_1 norm at the group level to promote group sparsity. A few different group structures that may be used to obtain specific sparsity patterns were explored by Jenatton et al. (2011), though the challenge of group construction was not discussed since it requires prior information about the problem under consideration and the required sparsity patterns. Since our work has a well-defined goal, i.e., the variable sparsity in a rule ensemble, we can develop a natural way to construct the groups. To do this, we first choose our set of groups $\mathcal{G} = \{G_1, \dots G_P\}$ such that each group G_p, corresponding to the variable x_p, contains the indices of the rules which involve the variable x_p. That is, if there are n_p number of rules containing variable x_p, then $G_p = \{p_1, \dots p_{n_p}\}$, where $\{p_1, \dots p_{n_p}\} \subseteq \{1, \dots K\}$ and all rules r_{p_j} $(1 \leq j \leq n_p)$ contains the variable x_p in at least

one of its combinatorial statements. Next, for each group G in $\mathcal{G}$, we define $\boldsymbol{\beta}_G$, a vector in $\mathbb{R}^{|G|}$, that consists of the elements of $\boldsymbol{\beta}$ belonging to G. Now, the formulation (6) could be further developed as

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} l(\boldsymbol{\beta}, \mathbf{x}) + \lambda_1 \cdot \|\boldsymbol{\beta}\|_1 + \lambda_2 \cdot \sum_{p=1}^{P} q_p \|\boldsymbol{\beta}_{G_p}\|_2. \tag{8}$$

Note that here q_p is an optional weight for each group. In our case, $q_p = \sqrt{1/|G_p|}$ to normalize the penalty term for groups of varying sizes. The hyperparameters λ_1 and λ_2 can be determined by cross-validation. The obtained solution is such that the potential nonzero patterns in the model are a complement of an intersection of a subset of groups. Fig 3 provides a representation of the sparsity structure that this regularization will induce. For example, if the group 4 is left out of the model, then all the coefficients belonging to this group will be zero. Since group 4 corresponds to the fourth variable (x_4), and contains coefficients of all rules containing x_4, all such rules are left out of the model. Thus, we are left with a complement of the intersection of the groups with group 4, and obtain a subset of rules that does not contain variable 4. The ℓ_1 regularization term here produces general within-group sparsity among all rules to select the most significant rules. Therefore

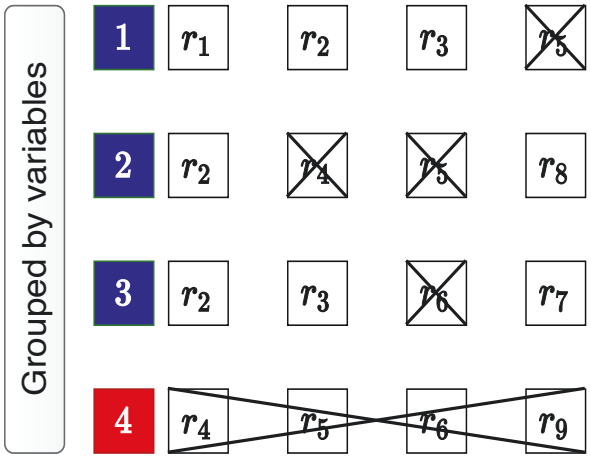


Figure 3 **Example of variable-sparse structure induced by overlapping group lasso regularization in SURVFIT. When the group corresponding to variable 4 (red) is left out, it zeroes out all the the coefficients of all rules containing variable 4 (crossed out).**

a larger value of λ_2 will induce greater group level sparsity in the resulting coefficients, $\hat{\boldsymbol{\beta}}$, which, for the selection of groups $\mathcal{G}$ proposed by us will mean greater degree of variable sparsity in the extracted rule set.

2.4. Optimization Strategy

The challenge to solve the formulation (8) is that the overlapping group structure introduces interdependency among the decision variables of the optimization problem. For non-overlapping groups (Yuan and Lin 2006), efficient algorithms that depend on seperability of variables, such as block coordinate descent can be applied (Simon et al. 2013b). To overcome this challenge, we reformulate this problem as a second-order cone program (SOCP) in Section 2.4.1, and use the interior point method to solve it optimally. However, this strategy will increase the problem size and therefore the computational cost. Alternative methods that may be more efficient include proximal methods, where a key challenge is to develop efficient solutions of the proximal operator. For instance, Chen et al. (2012) proposed a smoothing proximal gradient method where a smooth approximation of the overlapping group lasso norm (7) and the gradient of this approximation are derived. This smoothing strategy enables a fast iterative shrinkage-thresholding algorithm (FISTA) (Beck and Teboulle 2009) to solve the problem. Lei Yuan et al. (2013) proposed the FOGLASSO algorithm using an approximate dual of the proximal operator of the overlapping group lasso norm and its solution. We discuss their algorithm briefly in the following section. Chen et al. (2012) showed that the objective function of their semi-smooth approximation converges to the optimal solution, and Lei Yuan et al. (2013) observed that though FOGLASSO lacks a convergence guarantee, their algorithm almost always converges to the optimal solution in practice. However, neither study the sparsity structure of the solution they obtained in comparison to the sparsity structure of a solution which does not employ approximations. An understanding of the solution structure obtained is critical in high dimensional problems with multiple optimal solutions where the goal is not just to obtain a solution with an optimal value but also to minimize the number of nonzero coefficients in the solution.

2.4.1. SOCP optimization. In what follows, we cast the formulation (8) as a second-order cone program (SOCP). First, we introduce a variable, $z_i = \min(0, t(\mathbf{x}_i) - t_i), \forall i \in \{i \mid \delta_i = 0\}$. Then, we linearize the first regularization term in (8) by introducing two new variables, $\boldsymbol{\beta}^+, \boldsymbol{\beta}^-$ such that $\boldsymbol{\beta} = \boldsymbol{\beta}^+ - \boldsymbol{\beta}^-$, $|\boldsymbol{\beta}| = \boldsymbol{\beta}^+ + \boldsymbol{\beta}^-$, and $\boldsymbol{\beta}^+, \boldsymbol{\beta}^- \geq 0$. Third, to deal with the ℓ_2 overlapping group norm, we introduce new variables $\mathbf{y} \in \mathbb{R}^P$ such that

$$\mathbf{y}_p \geq \|\boldsymbol{\beta}_{G_p}\|_2 \quad \forall p \in \{1, \dots P\}. \tag{9}$$

The reformulated problem can be written as

$$\min_{\boldsymbol{\beta}^+,\boldsymbol{\beta}^-,\mathbf{z},\mathbf{y}} \ \|\mathbf{A}(\boldsymbol{\beta}^+ - \boldsymbol{\beta}^-) - \mathbf{t}\|_2^2 + \gamma\|\mathbf{z}\|_2^2 + \lambda_1 \cdot (\boldsymbol{\beta}^+ + \boldsymbol{\beta}^-) + \lambda_2 \sum_{p=1}^{P} q_p \mathbf{y}_p, \tag{10}$$

with the constraints,

$$\boldsymbol{\beta}^+, \boldsymbol{\beta}^- \geq 0, \tag{11}$$

$$\mathbf{z} \leq 0, \tag{12}$$

$$\mathbf{z} \leq \mathbf{B}(\boldsymbol{\beta}^+ - \boldsymbol{\beta}^-) - \mathbf{c}, \tag{13}$$

$$\mathbf{y}_i \geq \|(\boldsymbol{\beta}^+_{G_i} - \boldsymbol{\beta}^-_{G_i})\|_2 \quad \forall i \in \{1, \dots P\}. \tag{14}$$

Here $\mathbf{A}$ is the rule data matrix corresponding to event observations, $\mathbf{B}$ is the rule data matrix corresponding to censored observations and, $\mathbf{t}$ and $\mathbf{c}$ are the event times and censor times respectively, and (14) is a second-order cone constraint. This problem can now be solved by standard SOCP solvers. We use CPLEX in our experiments.

2.4.2. FOGLASSO. Another algorithm that can solve the formulation approximately is the FOGLASSO algorithm (Lei Yuan et al. 2013) implemented in the **SLEP** (Liu et al. 2009) MATLAB package. Here we briefly introduce some details of the FOGLASS algorithm for the sake of completeness. FOGLASSO uses an accelerated proximal gradient method where the coefficient estimates at each iteration are, $\boldsymbol{\beta}_{i+1} = \pi^{\lambda_1/L}_{\lambda_2/L}(s_i - \frac{1}{L} l'(s_i))$, where s_i is the affine combination of the current and previous estimates $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_{i-1}$ as used in FISTA (Beck and Teboulle 2009), L_i is an appropriate constant determined via backtracking line search by the Armijo-Goldstein condition (Armijo 1966), and $\pi(\cdot)$ is the proximal operator of the non-smooth regularization term, $\phi^{\lambda_2}_{\lambda_1}(\boldsymbol{\beta}) = \lambda_1 \cdot \|\boldsymbol{\beta}\|_1 + \lambda_2 \cdot \sum_{p=1}^{P} q_p \|\boldsymbol{\beta}_{G_p}\|_2$. Then, the main optimization problem of the proximal operator is derived to be

$$\pi^{\lambda_2}_{\lambda_1}(\mathbf{v}) = \arg\min_{\boldsymbol{\beta} \in \mathbf{R}^P} \Big\{ g^{\lambda_2}_{\lambda_1}(\boldsymbol{\beta}) = \frac{1}{2}\|\boldsymbol{\beta} - v\|^2 + \phi^{\lambda_2}_{\lambda_1}(\boldsymbol{\beta}) \Big\}. \tag{15}$$

FOGLASSO uses an efficient computational algorithm to solve this proximal operator by reformulating it as a smooth and convex dual problem. A pre-processing step is developed to identify many zero groups, which reduces the complexity of the optimization problem. However, a trade off is that their proximal operator solution is inexact, and it is stated that the optimal convergence rate is not guaranteed, though the algorithm works well in practice.

2.4.3. Scalability and Computational Analysis Here we provide an in-depth analysis of the computational complexity of SURVFIT as well as the existing baseline methods in the literature.

The package **pre** uses the Cox-lasso formulation to extract rules. The time complexity for Cox-regression is $\mathcal{O}(NK^2)$ while that of the Cox-lasso is $\mathcal{O}(NK)$ (Wang et al. 2019). However, there are significant computational challenges to optimizing the Cox partial likelihood loss function. The partial likelihood does not naturally decouple over individuals or subsets of individuals, therefore when regularized with a non-smooth term such as the overlapping group lasso, first order proximal gradient descent methods like FOGLASSO, prox-Grad and alternating direction method of multipliers (Boyd et al. 2004) cannot be used. This also means that the stochastic gradient-based optimization methods are not suitable for the task (Tarkhan and Simon 2020). We have not come across any works that have addressed these computational challenges in our review. One approach could be to use the standard Newton-Ralphson second order scheme (Boyd et al. 2004), however this is not practical for even medium-sized problems because it involves inverting large matrices at each iteration, which itself has complexity $\mathcal{O}(K^3)$, infeasible to solve our problem of rule selection since we start with a large number of rules. Our approach of using a quadratic loss function instead of Cox-partial log likelihood loss skirts this issue, allowing us to use efficient first order and second order optimization schemes (8).

In this paper, we have discussed two different optimization schemes, SOCP and FOGLASSO to solve Problem (8). Using the quadratic loss function instead of the Cox partial likelihood loss allows us to formulate the problem as a second order cone program and solve it using CPLEX solver. The solver uses the barrier interior point method (IPM), a second order method known to converge in $\mathcal{O}(\log(\frac{1}{\epsilon}))$ iterations, where ϵ is the accuracy at convergence. The per-iteration time complexity of IPM (Boyd et al. 2004) in our case is $\mathcal{O}\big(K^2(N+\sum_{p\in 1\cdots P}|G_p|)\big)$. First order proximal gradient methods such as FOGLASSO (Liu et al. 2009) and Prox-Grad (Chen et al. 2012) have been proposed to efficiently solve the problem of convex loss functions regularized with an overlapping group lasso loss. The algorithms take more iterations to converge than the second-order IPM algorithm, i.e $\mathcal{O}(\frac{1}{\epsilon})$ iterations. However, the per iteration complexity is lower by orders of magnitude thereby reaching convergence faster. For Prox-Grad, the per-iteration complexity is $\mathcal{O}(NK+\sum_{p\in 1\cdots P}|G_p|)$ (Chen et al. 2012). In practice IPM is found to be more accurate

than proximal methods though proximal methods are more efficient and scalable for large-scale problems. Thereby, we see that applying an overlapping group regularization term to induce variable sparsity for the standard Cox partial likelihood loss has a time complexity of atleast $\mathcal{O}(NK^3)$ while our formulation can be solved by first order methods in $\mathcal{O}(NK + \sum_{p\in 1\cdots P}|G_p|)$ and second order methods in $\mathcal{O}\big(K^2(N + \sum_{p\in 1\cdots P}|G_p|)\big)$. This is much more efficient in high-dimensional problems such as rule extraction where often $K >> N$, i.e., the number of rules is much greater than the number of observations.

3. Interpretability of the Rules

Most rule learning methods in the literature of machine learning only concern rule discovery, namely the identification of important rules from data such as Rulefit (Friedman and Popescu 2008) and **pre** (Fokkema 2017). For instance, a rule suggests an interaction among a set of variables, and the essence of an interaction is that the variables give a greater effect when combined than taken individually. A rule learning algorithm may generate a set of rules, but it cannot prove that the interactions are genuine. Here, we further develop a rule analysis framework that employs a combination of statistical methods such as survival data analysis, hypothesis testing, and regression analysis to evaluate the significance and to better understand the implications of the discovered rules in various contexts.

3.1. Statistical Testing.

We use statistical testing to evaluate whether the extracted rules are significantly associated with the outcome. Specifically, we analyze whether the subjects endorsing each rule have a significantly higher or lower risk of onset of the event as compared to subjects not endorsing the rule. For this goal, the Kaplan-Meier curve is used to study the separation of the survival functions of the groups of observations defined by each rule. We also employ the log-rank test (Mantel 1966), which is a hypothesis testing method used to examine differences in risk of event occurrence between the two groups.

3.2. Decomposition Analysis.

Decomposition analysis is used to examine the rules to see if the interactions among the variables are genuine. Basically, we decompose the rule by removing one variable at a time and evaluating the impact of this removal, i.e., by statistically evaluating the difference using the Kaplan-Meier curve and log-rank tests. If the removal of a variable is found to have little impact on the overall significance of the rule, which is possible since the

rule learning algorithm uses a greedy predictive metric to guide the rule selection process, then we should trim the rule by removing this variable. On the other hand, the variable which has the greatest impact on the significance of the rule is said to be the dominant or critical variable of that rule. And if the combination of two or more variables has a much higher association with the risk than either of the variables taken individually, then the interaction of those two variables is highly significant in predicting the risk. Those are possible scenarios the decomposition analysis could shed light on and reveal more understanding of the rules and their constituent variables.

3.3. Sensitivity Analysis.

Besides the combinatorial characteristic of the rules, cutoff values of the constituent variables are also essential information in defining the interactions among the variables. Sensitivity analysis is conducted to evaluate how sensitively the statistical significance of the rule depends on the cutoff values of the variables used in the rule. We study the change in the odds ratio of the two groups defined by the rule as the cutoff value of a variable in a rule changes. The odds ratio (OR) (Bland and Altman 2000) is the ratio of the probability of event occurring in the subgroup endorsing the rule to the ratio of the event occurring in the other subgroup. This sensitivity analysis would reveal different scenarios for the cutoff values as well, e.g., there is sometimes indeed a best cutoff value for a factor, with a cutoff value that maximizes the statistical significance of the rule, and around the best value there is a either a sharp or smooth descending slope. For some other variables, there seems to be a range of cutoff values that are equally good. Thus, sensitivity analysis could reveal unique insights regarding the variables and the rules that engage them.

4. Numerical Experiments

In this section we conduct numerical experiments to evaluate the proposed SURVFIT method, compare the solutions produced by the two optimization strategies described in Section 2.4, and compare SURVFIT with the baseline approach that uses a regularized cox regression model and survival random forest. One of our goals is to demonstrate the variable sparsity property of SURVFIT. Therefore, the rules extracted with (8) and without (6) doubly sparse regularization, and their decomposition, and sensitivity analysis, are also presented. Predictive and variable selection performance for each of the models are evaluated over 100 repetitions of the same experimental setup, with 80/20 partitions of the data for training/testing.

4.1. Evaluation Criteria

The following measures are used to rank, and evaluate the significance of the rules extracted by SURVFIT.

1. Importance. We rank the rules obtained by our model by the importance measure, I(r) which is defined as

$$\mathrm{I}(r) = \hat{\beta}_r \sqrt{s(r)(1-s(r))}, \tag{16}$$

where $s(r)$ is the support and $\hat{\beta}_r$ is the coefficient estimate of rule r.

2. Support. The support of a rule, $s(r)$ is defined as the fraction of total observations that endorse the rule,

$$s(r) = \frac{\sum_{i=1}^{N} r(\mathbf{x}_i)}{N}. \tag{17}$$

3. Odds Ratio. The odds ratio is the ratio of the odds of event occurence in the data endorsing the rule to the odds of event occurence in data not endorsing the rule.

Furthermore, the following measures are used to evaluate and compare the performance of our method with baseline models.

1. Concordance index (C-Index): Harell's concordance index (Harrell et al. 1982) is used to estimate and compare the predictive performance of survival models. The c-index estimates the probability that in randomly selected pair of test subjects, the subject with the earlier event occurence has an earlier model prediction of event time. Therefore a completely random prediction will achieve a c-index of 0.5.

2. False positive rate (FPR): The FPR is a measure of the models capability to select a sparse set of significant variables. It is defined as the ratio of the number of incorrect (or noisy) variables selected by the model to the total number of variables.

4.2. Synthetic Data

We simulate a dataset consisting of $N = 2000$ observations with $P = 60$ variables using an approach similar to the one adopted in Friedman and Popescu (2008). The event times (18) are simulated such that only 7 variables ($x_1 - x_7$) affect the target, another 7 variables ($x_{18} - x_{14}$) are correlated in varying degrees to the first 7 variables, and the remaining 46 variables are purely noise. The response variable t_i for each input $\mathbf{x}_i$ is taken to be

$$t_i = F(\mathbf{x}_i) + \epsilon_i, \tag{18}$$

where the function $F(x)$ is defined as

$$F(x) = 4\prod_{j=1}^{3}(-(1-x_j)^2) - 0.55 * \exp\{-2(x_4 - x_5)\} + 1.75 * \sin(x_6 - x_7) \tag{19}$$

and $\epsilon \sim N(0, \sigma^2)$, where σ^2 is chosen to keep a signal to noise ratio of 3. The response $\mathbf{t}$ is then scaled to make sure it is positive. The parameters of this function are chosen in a way to obtain approximately equal representation of each predictive variable in the exhaustive rule list. We assume that 35% of the simulated data is censored. To account for this, the event times of a random subset consisting of 35% of the data are multiplied with a uniform random variable to simulate the censored times. The remaining 65% of the data is assumed to be complete. To simulate each of the correlated variables in $\{x_8 \dots x_{14}\}$, we first sample a correlation $\rho \in (0, 1)$ from a uniform distribution by which it is correlated to the corresponding original variable in $\{x_1 \dots x_7\}$, then we choose $x = \rho \cdot \sigma_{x^*} x + \sqrt{1-\rho^2} \cdot \sigma_x x^*$, where x is the correlated variable in $\{x_8 \dots x_{14}\}$ and x^* is the residual of a least-squared regression between x and its corresponding original variable in $\{x_1 \dots x_7\}$. The response simulation model (19) involves explicit interactions between (x_1, x_2, x_3) in the first term, (x_4, x_5) in the second term, and between (x_6, x_7) in the third term.

4.3. Synthetic Data Results

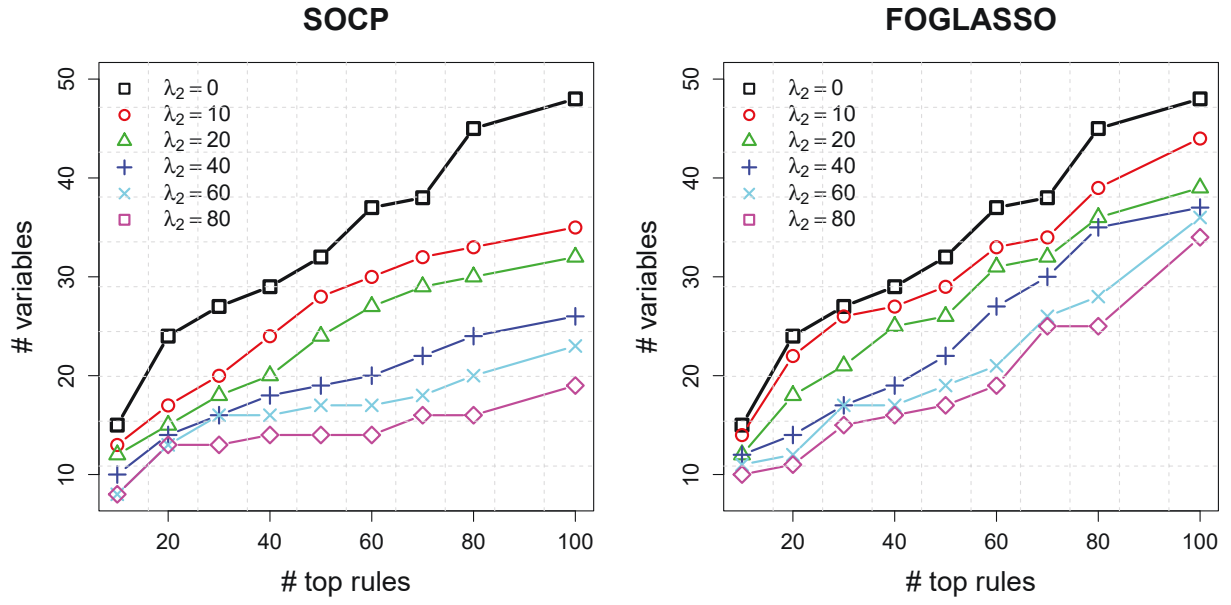


Figure 4 **Number of variables included in the top rules extracted at various values of λ_2 for the SOCP (left) and FOGLASSO (right) optimization methods.**

4.3.1. Comparison of the SOCP algorithm with FOGLASSO. Fig. 4 compares the number of variables included in the top ranked rules extracted at various values of λ_2 for the SOCP and FOGLASSO methods. Firstly, both algorithms show the effectiveness of the SURVFIT formulation to produce a doubly-sparse set of rules which are also sparse in the number of variables involved in the rules, i.e., it is observed that the extracted rules contain fewer variables when λ_2 is increased. Secondly, we observe that comparing with our SOCP algorithm, FOGLASSO leads to rules with more variables. It is worth mentioning that in our experiments, solving the problem using proximal smoothing algorithm introduced by Chen et al. (2012) leads to a selection of rules which had even more variables than FOGLASSO. This difference in the structure of the solutions under these algorithms may be attributed to the approximate nature of the algorithm used to solve the proximal operator in FOGLASSO, and the smoothing approximation of the overlapping group norm (7) used by Chen et al. (2012) which lead to a less sparse solution in terms of groups i.e variable selection, a trade off for their computational advantage.

4.3.2. Quality and efficiency of rule extraction. The hyperparameters, $\lambda_1 = 5$ and $\lambda_2 = 20$ are chosen after cross validation for the following analysis. The SURVFIT algorithm then extracts the highly predictive rules which are also sparse in the number of variables among them. SURVFIT, like other rule models based on sparse regularization, produces a rank for each rule. One way to compare different methods is to see the quality of highly ranked rules, i.e if noisy or low quality rules have high rank, the algorithm is less useful. The rules are ranked based on the *importance* measure introduced in Section 4.1. In our experiment, we compare the top 8 rules which are extracted by SURVFIT (in Table 2) and the rules extracted by ℓ_1 penalized optimization without the variable sparsity penalty (in Table 1). One main interest of this numerical study is to see if SURVFIT could detect significant rules without falsely introducing noisy variables. Table 2 presents the top 8 rules extracted via SURVFIT which involve 7 variables, x_1, x_2, $x_4 - x_7$ and a singular false positive, x_{11}. (The significant variable x_3 is not picked in the top 8 rules though it does show up in the total list of significant rules with non-zero coefficients). This is consistent with the ground truth that has been used in the synthetic data generation as shown in (18). A remarkable observation is that SURVFIT is resilient to the noise in the data which was designed for variables $x_8 - x_{14}$ to be statistically correlated with $x_1 - x_7$. To further demonstrate this point, Table 1 lists the top 8 rules extracted without

doubly-sparse regularization, i.e., only employing ℓ_1 norm penalty to obtain sparsity in the cardinality of rules. In contrast to SURVFIT we can observe that the rules in Table 1 contain a total of 15 variables, of which many are false positives. Thus, doubly-sparse regularization used in SURVFIT is effective to recover the true variables and genuine interactions in the data generating process. Fig. 4 shows the number of variables involved in the top rules returned by SURVFIT for different values of λ_2. Increasing the value of the variable-sparsity parameter, λ_2 leads to rules which are more sparse in the number of variables. For instance, at $\lambda_2 = 0$, the top 10 rules involve 16 variables, while for $\lambda_2 = 80$, the top 10 rules involve only 7 variables. The difference is even more stark when comparing the top 100 rules: there are close to 50 variables at $\lambda_2 = 0$, while at $\lambda_2 = 80$, the top 100 rules only contain 16 variables. We conclude that our proposed SURVFIT approach is able to extract a doubly-sparse set of rules involving only the most significant variables in the data.

4.3.3. Rule analysis. Table 2 not only presents the top 8 rules extracted via SURVFIT, but also their decompositions, p values of the log-rank test, and support (4.1) of each of these rules. Also, Fig. 5 shows the Kaplan-Meier survival curves for these rules. We observe from the Kaplan-Meier curves that, endorsement of rules 1-6 and rule 8 is associated with higher survival rates, while endorsement of rule 7 is associated with lower survival rate. As we show in the following discussion, these rules recover the variable effects and interactions on mortality encoded in the data generation process. The decomposition analysis (besides the p-values shown in Table 1, the decomposition curves for the rules are also plotted in Fig. 6) also reveals interesting insights into the data. For instance, on rule 2 we can see that there is a real interaction between x_4 and x_5, as removal of either hugely impacts the significance of the rule. This is consistent with the ground truth model as the way the two variables are incorporated into the data-generating mechanism is through the functional $\exp\{-2(x_4 - x_5)\}$. A similar observation holds true for rule 3 and its decomposition analysis reveals the interaction between x_6 and x_7. The decomposition analysis on rule 4 is also insightful, as it actually shows that with x_5 alone the significance of the rule is stronger, indicating that there is no synergistic interaction between x_1 and x_5 so there is no need to keep variable x_1 in the rule. This shows that rule 4 contains noise resulting from the greedy nature of the tree-growing algorithm used. While the double-sparsity enforced by SURVFIT aims to reduce this noise, we also need decomposition analysis to further filter

out any residual noise. Similar conclusion holds true for the decomposition analysis on rule 5, rule 6, and rule 7. We could still keep some significant, but not explicitly interacting variables in the rules, e.g., x_2 in rule 5 and x_1 in rule 6, since it is hard to call these two rules interrupted by noise, given that x_2 and x_1 contribute to the overall significance of the rules though their contributions are quite marginal. Here there is no obvious contradiction with the ground truth model since all the variables are truly involved in the ground truth, although not in explicit interaction terms, an implicit interaction exists.

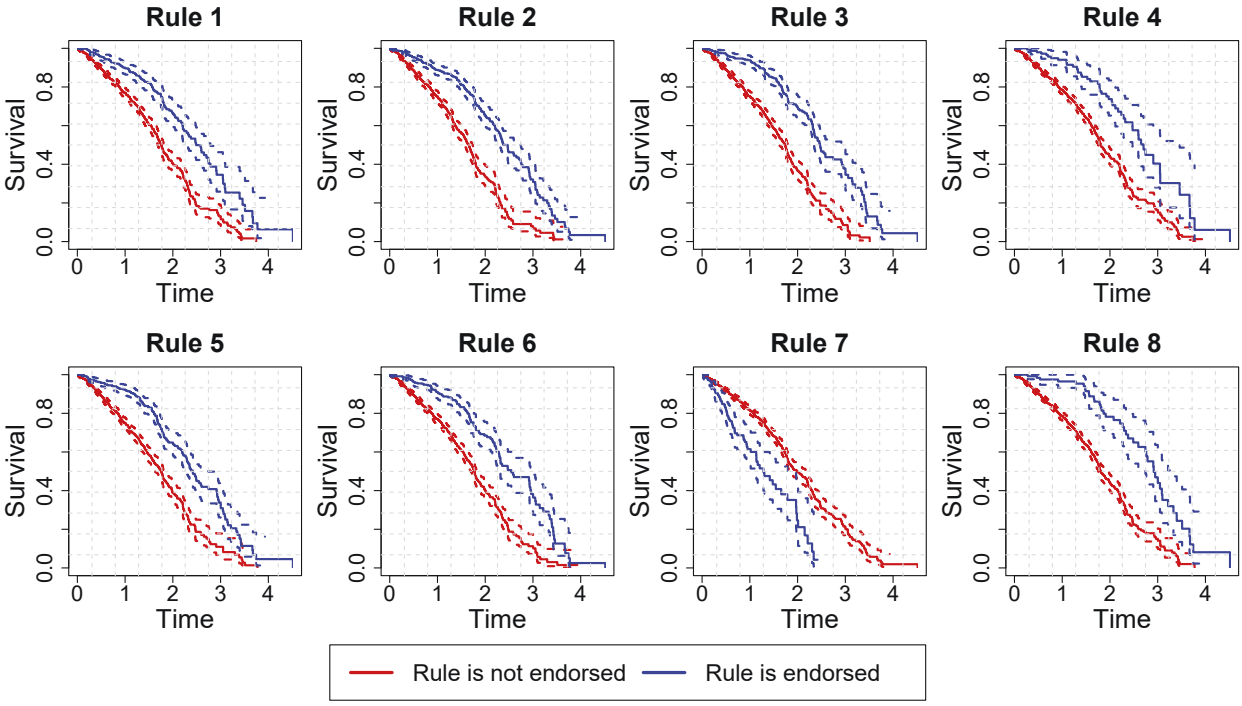


Figure 5 **Kaplan–Meier survival curves with 95% confidence intervals for each rule in Table 2**

Next we analyze the sensitivity of the cutoff values of the variables. Fig. 7 shows how the overall significance of the rules, as measured by the odds ratio and its 95% confidence interval changes with change in the cutoff value of the variables in the rules. For instance, in the sensitivity analysis of Rule 1, we see that increasing the cutoff value of x_5 reduces the odds ratio, meaning that at higher values of x_5 the probability of event occurence falls off. This is consistent with both the Kaplan Meier curve associated with rule 1 as well as the data-generating mechanism. The sensitivity analysis of rule 2 shows how the odds ratio would change with change in cutoff of each of x_4 and x_5, while keeping the other cutoff constant. At smaller cutoff values of x_4, the odds ratio is greater than 1 signifying higher

Table 1 **Top 8 Rules Identified without double sparsity penalty (6) and Corresponding Log-Rank p-Values**

Rule ID	Rule	p Value	Support (%)
1	$x_4 > 0.532$ AND $x_{49} > 0.033$	$8e-13$	37.05
1a	$x_4 > 0.532$	$4e-15$	38.3
1b	$x_{49} > 0.033$	0.01	96.6
2	$x_5 > 0.47$ AND $x_{10} <= 0.705$	$3e-19$	42.25
2a	$x_5 > 0.47$	$3e-16$	46
2b	$x_{10} <= 0.705$	$3e-04$	92
3	$x_2 <= 0.309$ AND $x_5 > 0.5$ AND $x_{11} <= 0.289$	$6e-12$	6.40
3a	$x_5 > 0.5$ AND $x_{11} <= 0.289$	$2e-14$	18.40
3b	$x_2 <= 0.309$ AND $x_{11} <= 0.289$	$2e-12$	15.90
3c	$x_2 <= 0.309$ AND $x_5 > 0.5$	$1e-13$	14.90
4	$x_2 <= 0.45$ AND $x_6 > 0.486$ AND $x_{57} <= 0.892$	$3.5e-18$	23.25
4a	$x_6 > 0.486$ AND $x_{57} <= 0.892$	$2.5e-17$	44.05
4b	$x_2 <= 0.45$ AND $x_{57} <= 0.892$	$5e-12$	50.60
4c	$x_2 <= 0.45$ AND $x_6 > 0.486$	$2e-18$	23.55
5	$x_1 <= 0.25$ AND $x_6 > 0.442$ AND $x_{11} <= 0.346$	$6e-13$	6.90
5a	$x_6 > 0.442$ AND $x_{11} <= 0.346$	$6e-16$	25.80
5b	$x_1 <= 0.25$ AND $x_{11} <= 0.346$	$1e-09$	15.25
5c	$x_1 <= 0.25$ AND $x_6 > 0.442$	$1e-13$	13.45
6	$x_6 > 0.48$ AND $x_7 > 0.48$ AND $x_{30} <= 0.893$	$2e-22$	20.85
6a	$x_7 > 0.48$ AND $x_{30} <= 0.893$	$3e-10$	46.20
6b	$x_6 > 0.48$ AND $x_{30} <= 0.893$	$2e-17$	44.90
6c	$x_6 > 0.48$ AND $x_7 > 0.48$	$3e-22$	21.00
7	$x_4 <= 0.494$ AND $x_9 <= 0.553$ AND $x_{13} > 0.549$	$3e-12$	10.35
7a	$x_9 <= 0.553$ AND $x_{13} > 0.549$	$2e-11$	18.60
7b	$x_4 <= 0.494$ AND $x_{13} > 0.549$	$9e-09$	12.60
7c	$x_4 <= 0.494$ AND $x_9 <= 0.553$	$4e-15$	44.30
8	$x_4 <= 0.323$ AND $x_{24} > 0.886$ AND $x_{28} > 0.0043$	$3e-10$	61.45
8a	$x_{24} > 0.886$ AND $x_{28} > 0.0043$	0.2	98.00
8b	$x_4 <= 0.323$ AND $x_{28} > 0.0043$	$2e-11$	62.65
8c	$x_4 <= 0.323$ AND $x_{24} > 0.886$	$4e-15$	61.60

Table 2 Top 8 Rules Identified with double sparsity penalty (8) and Corresponding Log-Rank p-Values. The final rules selected after decomposition analysis are highlighted in gray.

Rule ID	Rule	p Value	Support (%)
1	$x_5 > 0.687$	$7.5e-15$	21.5
2	$x_4 <= 0.52$ AND $x_5 > 0.395$	$7e-25$	32.80
2a	$x_4 <= 0.52$	$8e-15$	60.20
2b	$x_5 > 0.395$	$3e-14$	54.40
3	$x_6 > 0.479$ AND $x_7 > 0.48$	$3e-23$	21.0
3a	$x_6 > 0.479$	$2e-17$	45.25
3b	$x_7 > 0.48$	$3e-10$	46.55
4	$x_1 < 0.4$ AND $x_5 > 0.71$	$2e-9$	8.05
4a	$x_1 < 0.4$	$2e-6$	44.80
4b	$x_5 > 0.71$	$1e-10$	18.25
5	$x_2 <= 0.45$ AND $x_6 > 0.485$	$6e-18$	23.55
5a	$x_2 <= 0.45$	$4e-10$	51.00
5b	$x_6 > 0.485$	$2e-17$	44.55
6	$x_1 <= 0.52$ AND $x_6 > 0.625$	$2e-18$	18.20
6a	$x_1 <= 0.52$	$4e-4$	58.10
6b	$x_6 > 0.625$	$5e-17$	30.75
7	$x_1 >= 0.33$ AND $x_2 > 0.65$ AND $x_7 > 0.428$	$4e-11$	9.30
7a	$x_2 > 0.65$ AND $x_7 > 0.428$	$2e-04$	14.95
7b	$x_1 >= 0.33$ AND $x_7 > 0.428$	0.9	33.90
7c	$x_1 >= 0.33$ AND $x_2 > 0.65$	$2e-12$	17.35
8	$x_2 < 0.31$ AND $x_5 > 0.5$ AND $x_{11} <= 0.288$	$6e-12$	6.4
8a	$x_5 > 0.5$ AND $x_{11} <= 0.288$	$2e-14$	18.40
8b	$x_2 < 0.31$ AND $x_{11} <= 0.288$	$2e-12$	15.90
8c	$x_2 < 0.31$ AND $x_5 > 0.5$	$1e-13$	14.90

risk levels while at higher cutoffs of x_5, the cutoff is lower than 1 signifying lower risk. The is again consistent with the ground truth since the time of event occurence depends on $-\exp(-2(x_4 - x_5))$, and a larger value of x_4 would decrease and a greater value of x_5 would increase the value of this term and hence decrease and increase the time of event

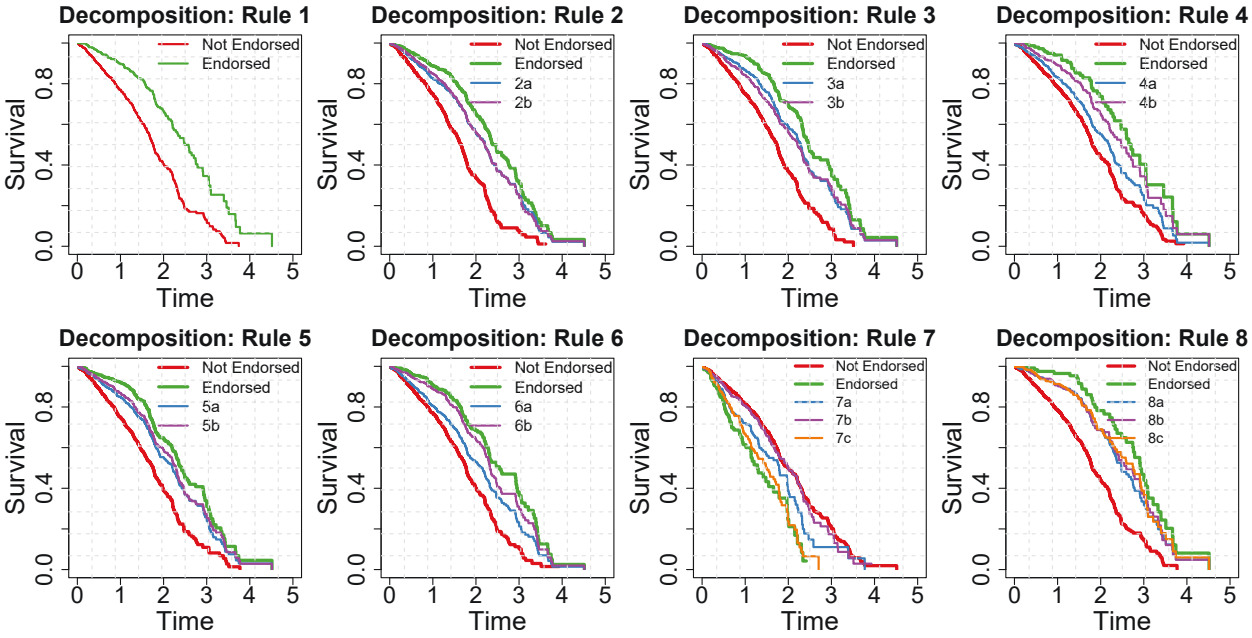


Figure 6 **Decomposition analysis for each rule in Table 2**

occurence, respectively. The sensitivity analysis of the other rules can also be interpreted in this context and seen to be consistent with our knowledge of the ground truth. Thus, the sensitivity analysis helps us understand how each variable affects the event risk under different conditions.

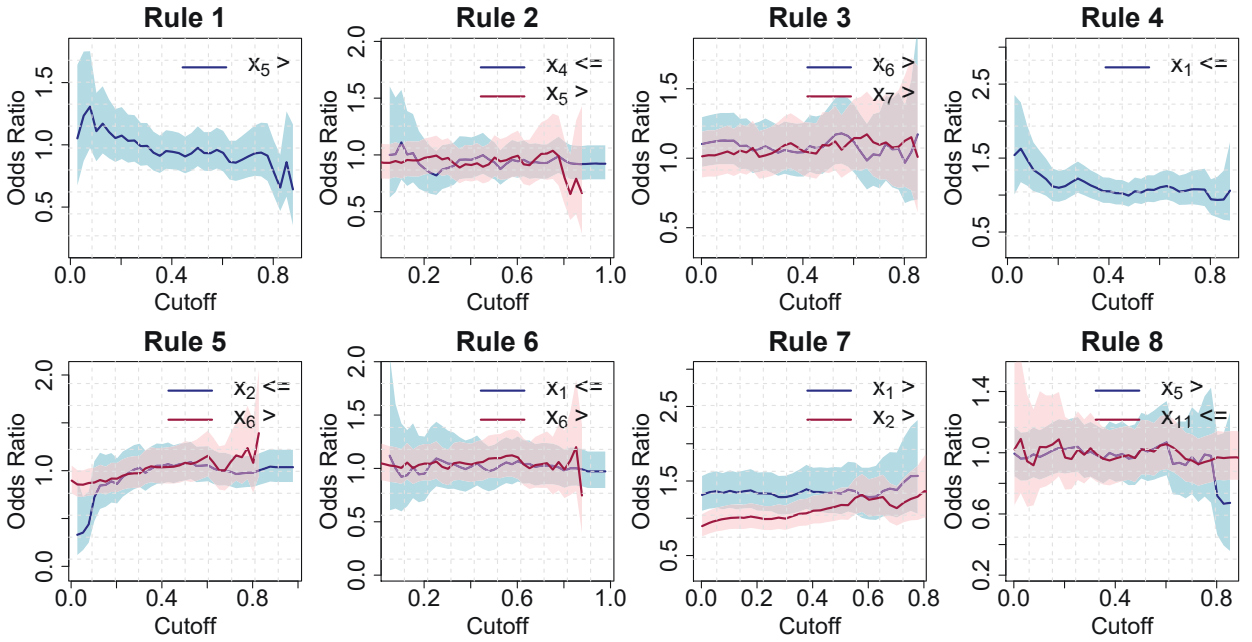


Figure 7 **Sensitivity analysis of rules from Table 2**

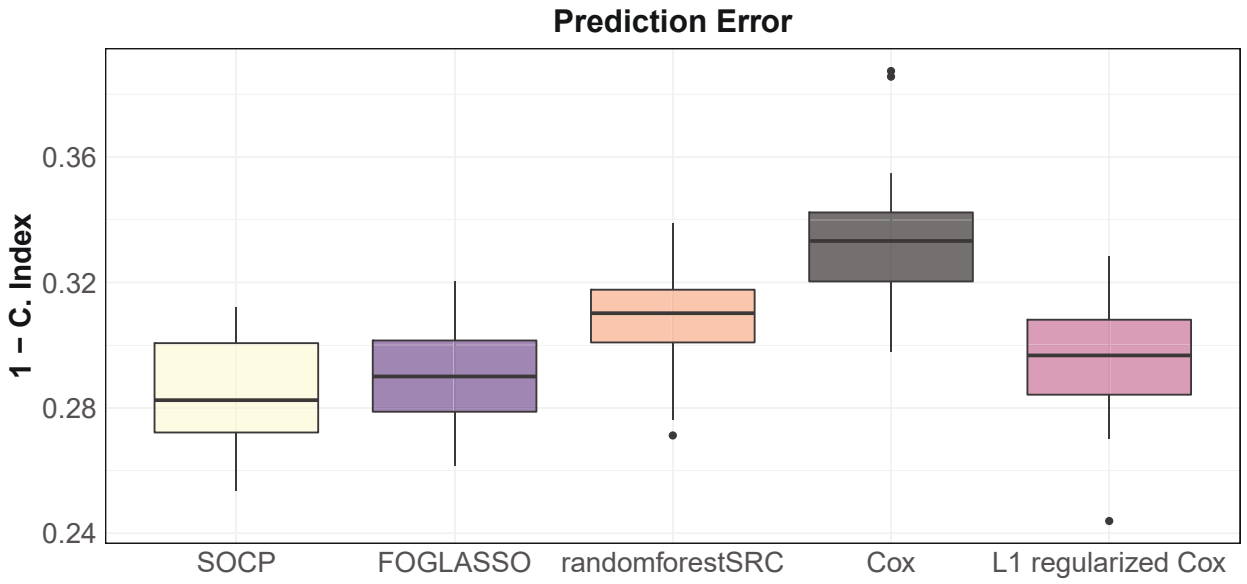


Figure 8 **Comparison of prediction error of SURVFIT with standard survival analysis methods on synthetic data**

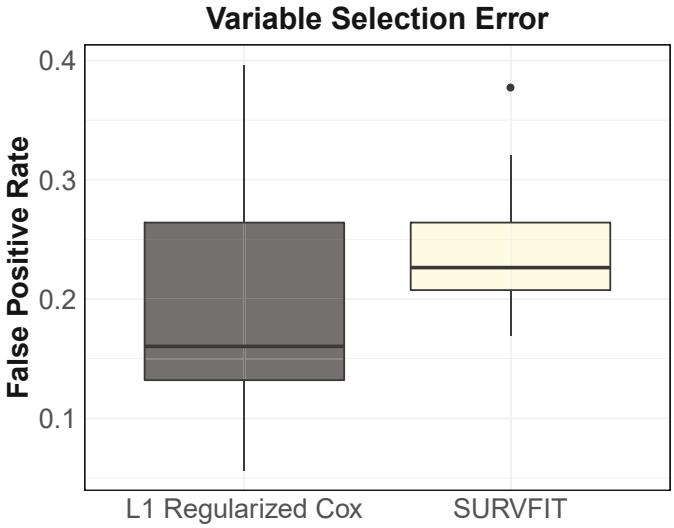


Figure 9 **Comparison of sparsity performance of Regularized Cox and SURVFIT model on synthetic data**

4.3.4. Comparison of predictive and sparsity performance of SURVFIT with standard survival models. We compare the predictive performance of the SURVFIT algorithm to standard survival analysis methods such as random survival forests implemented in **randomForestSRC** (Ishwaran et al. 2008), Cox regression (Cox 1972), and regularized Cox regression (Tibshirani 1997) using the concordance index (C Index) metric (4.1). The C Index for SURVFIT is calculated through the times-of-event occurence estimated by SURVFIT as shown in Section 4.1. Fig. 8 shows box plots of the estimates of prediction

error (1 - C Index) obtained over 100 independently sampled replicates. The following performance evaluation procedure has been adopted: first we sample a training set of 1600 observations, and then an independent test set of 400 observations. The different models are then trained on the training dataset and the reported performance evaluation is based on on the test set. This procedure is repeated 100 times to obtain the C-Index estimates for each of the different methods. It can be observed that the prediction errors of SURVFIT, when considering rules with all non-zero coefficients is lower than prediction errors of other methods. To compare the error rates of the different methods, we use the paired Wilcoxon rank sum test on our C-Index estimates. For each pair of the methods, we perform the following test:

Null hypothesis: $\text{C-Index}_1 = \text{C-Index}_2$

Alternative hypothesis: $\text{C-Index}_1 \neq \text{C-Index}_2$

The p-Values from each of these pairwise tests are provided in Table 3. It can be seen that the performance of the methods are significantly different from each other.

Table 3 **p-value of pairwise Wilcoxon rank sum test on C-index obtained by each of these methods on synthetic data**

	FOGLASSO	randomForestSRC	Cox	L1 regularized Cox
SOCP	0.0002563	9.3e-09	1.86e-09	0.00113
FOGLASSO		1.86e-08	1.86e-09	0.04592
randomForestSRC			1.82e-06	1.3e-07
Cox				1.86e-09

Our goal is not only to get a model that is accurate in terms of prediction but also exhibits sparsity in the number of variables. To do this we compare the false positives in the variables involved in the SURVFIT model with variables involved in the regularized Cox regression model. Fig. 9 compares a box plot of the variable selection error (false positive rate, (2)) of the regularized Cox model and SURVFIT obtained over 100 independently sampled replicates. As the figure shows, regularized Cox-regression does slightly better on average than SURVFIT on our synthetic data in terms of variable selection, although the spread of error is higher. The other models like randomForestSRC and Cox regression use all variables in the data, therefore a comparison of variable selection with these models is

not meaningful. The true positive rate, i.e., the proportion of correct variables identified is equal to 1 for both models, i.e., both SURVFIT and regularized Cox regression select all of the significant variables.

5. A Real-World Case Study: MIMIC Sepsis Data

MIMIC-III (Medical Information Mart for Intensive Care) (Johnson et al. 2016) is a comprehensive database comprising anonymized information relating to patients admitted to the Beth Israel Deaconess Medical Center in Boston, MA between 2001 and 2012. The data consists of over 53,000 adult ICU admissions during this time period. In this paper, we utilize a subset of inpatient admissions which were diagnosed with at least one of sepsis, or severe sepsis, or septic shock, which are increasingly severe sepsis conditions. This subset has 2,840 samples in total. Sepsis is a common ailment caused by infections and characterized by whole body inflammation which accounts for 2% of hospitalizations and 25% of ICU bed utilization's in the United States. It is the second leading cause of death among ICU patients, the third leading cause of death worldwide, and the main cause of hospital mortality (Gotts and Matthay 2016, Liu et al. 2014). Understanding mortality risk from sepsis would be beneficial for physicians in selecting a more efficient management approach. Several recent studies have focused on predicting mortality risk based on variables related to predisposition (Moreno et al. 2008), pre-existing and co-morbid conditions (Ford et al. 2016), cytokines and immune system interleukin's (Andaluz-Ojeda et al. 2012), and gene expression analysis (Sweeney et al. 2018). A recent study on early sepsis detection by Shashikumar et al. (2017) has used heart rate and blood pressure dynamics data. As the mechanism of how these variables impact the mortality risk is known to be complex, we use SURVFIT to extract rules and study the interactions among the variables. Out of the 2,840 patient observations in our dataset, 1,097 (38.6%) are mortal event instances with a record of time of death and the remaining are censored with time of discharge as the censor time. We investigated 78 variables in our analysis, consisting of patient characteristics such as age, race, gender, weight, clinical history; physiological measurements such as respiratory rate, blood pressure, heart rate, oxygen saturation etc., and summary statistics of physiological measurements and laboratory test results such as blood urea nitrogen, creatinine, white blood cell count, and hemoglobin etc.

5.1. Sepsis Survival Results.

We choose $\gamma = 5 \times 10^{-6}$, $\lambda_1 = 50$ and $\lambda_2 = 10$ for our SURVFIT model through cross-validation. The top 8 rules extracted by SURVFIT are presented in Table 4, along with their p-values of the log-rank test, their support, and the results of the decomposition and sensitivitiy analysis. We obtain a total of 13 significant variables involved in the top 8 rules affecting survival risk. Aspartate-aminotransferase, oxygen saturation (O_2-sat.), Alanine-aminotransferase, arterial-pH, age, heart-rate, Alanine-aminotransferase(tests), diastolic BP (noninv-dia-BP), length of stay, systolic BP (noninv-sys-BP) are the variables associated with sepsis mortality risk. We quantitatively, and descriptively evaluate their interaction effects on mortality. Each of these rules is significant, i.e., as shown in the p values of the log-rank test. The Kaplan–Meier curves of the rules are shown in Fig. 10. The Kaplan–Meier curves reveal that the rules 1, 2, 4, 5 and 7 are risk-reducing rules, i.e., patients who endorse these rules have less risk of mortality, and the rules 3, 6 and 8 are risk-increasing rules. Based on the decomposition analysis of the rules, i.e., p-values shown in Table 4 and survival curves of decomposition analysis shown in Fig. 11, we are able to gain a greater understanding of the nature of the interactions of the variable in each of the rules. For example, in rule 1, while both Aspartate-aminotransferase(mean) and oxygen saturation, O_2-sat.(mean) are significant in predicting the risk, O_2-sat.(mean) is the critical factor due to its lower p value. In rule 2, the interaction between Alanine-aminotransferase (mean) and arterial-pH (mean) is significant in predicting the mortality. Decomposition analysis of rule 3 demonstrates an interaction of heart-rate(sd) and arterial-pH(mean) to be highly significant while interaction between age and arterial-pH is not. The removal of heart-rate(sd) from rule 4 reduces the rule discrimination ability the most, making it the critical factor. Likewise, in other rules we observe that diastolic blood pressure (noninv-dia-BP), systolic blood pressure (noninv-sys-BP) and length of stay (total LOS) also influence the mortality rate. In rule 8, decomposition analysis reveals that noninv-sys-BP(mean) is the critical factor of the rule while heart-rate(tests) has no contribution despite being involved in the rule. The literature studying sepsis mortality supports our results, as the variables covered in these rules as well as their cutoff values have been found to be significant in predicting the mortality associated with sepsis. For example, rule 1 suggests that higher saturated oxygen, (O_2-sat(mean) > 92.5) is associated with a lower mortality risk. Our findings are consistent with the results found by Leone et al. (2009) who reported a

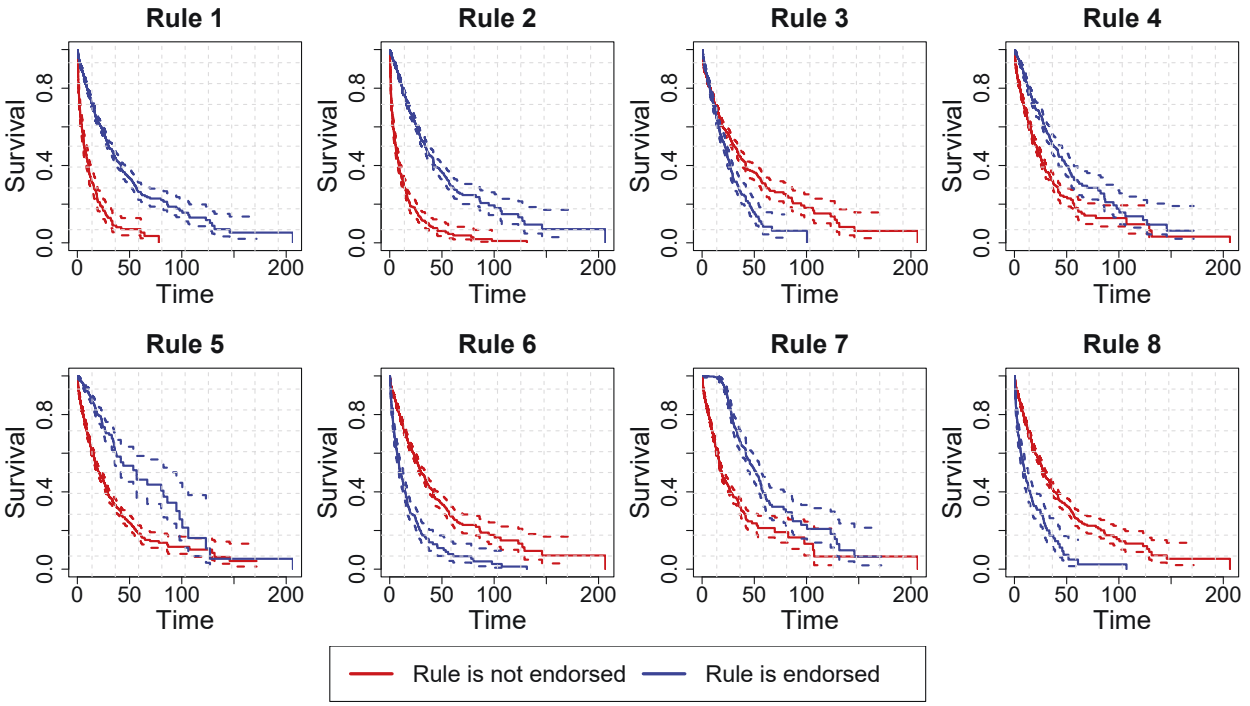


Figure 10 **Kaplan–Meier survival curves with 95% confidence intervals for rules in Table 4**

lower level of oxygen saturation in non-survivors as compared to survivors, and a value below 78 is associated with increased risk of mortality among patients of septic shock in their experiments. Alanine aminotransferase (alt), and aspartate aminotransferase (ast) are liver enzymes that are biomarkers of abnormal liver functions which is often found in sepsis patients (Nesseler et al. 2012). In rules 1 and 2 we see that a higher Aspartate-aminotransferase and Alanine aminotransferase signifies increased mortality risk, high levels of both enzymes have been found to significant predictors of sepsis associated liver injury (Dou et al. 2019, Zagory et al. 2017) in literature. In rule 6, standard deviation of oxygen saturation (O_2-sat(sd)) is found to be a significant predictor of mortality, and higher deviations are associated with higher mortality. A similar result was found by Krafft et al. (1993) who investigated spontaneous changes in oxygen saturated in sepsis patients, and reported a significantly higher number of severe changes in O_2-sat in non-surviving patients when compared to surviving patients. However, the literature does not discuss the significance of interactions between the variables found in our model.

Sensitivity analysis of some critical factors of the rules are reported in Fig. 12. The sensitivity analysis figures show the odds ratio (and 95% confidence interval) of the rules change when the cutoff values of variables are changed while keeping cutoffs of other variables in

Table 4 Top 8 rules identified with double sparse penalty from Sepsis survival data and their decomposition analysis. The final rules selected after decomposition analysis are highlighted in gray.

ID	Rule	p Value	Support
1	Aspartate-aminotransferase (mean)$<=308$ AND O_2-sat. (mean)>92.5	$2e-70$	89.6
1a	Aspartate-aminotransferase (mean)$<=308$	$2e-38$	91.62
1b	O_2-sat. (mean)>92.5	$2e-83$	97
2	Alanine-aminotransferase (mean)<2778.3 AND arterial-pH (mean)>7.2 AND O_2-sat. (sd)$<=3.23$	$8e-150$	81.83
2a	arterial-pH (mean)>7.2 AND O_2-sat. (sd)$<=3.23$	$3e-148$	81.93
2b	Alanine-aminotransferase (mean)<2778.3 AND O_2-sat. (sd)$<=3.23$	$8e-59$	86.97
2c	Alanine-aminotransferase (mean)<2778.3 AND arterial-pH (mean)>7.2	$8e-240$	92.18
3	age >73.85 AND heart-rate (sd) $<=38.74$ AND arterial-pH (mean) >7.25125	$3e-05$	37.21
3a	heart-rate (sd) $<=38.74$ AND arterial-pH (mean) >7.25125	$3e-240$	92.14
3b	age >73.85 AND arterial-pH (mean) >7.25125	$1e-05$	37.39
3c	age >73.85 AND heart-rate (sd) $<=38.74>7.25125$	$4e-19$	40.21
4	has.septicshock $=F$ AND Alanine-aminotransferase (tests) >3.5	$2e-23$	34.78
4a	has.septicshock $=F$	$3e-17$	50.38
4b	Alanine-aminotransferase (tests) >3.5	$1e-23$	69.82
5	noninv-dia-BP (mean)>34.3 AND O_2-sat. (sd)$<=5.8$ AND noninv-sys-BP (mean) >111.5	$4e-32$	25.03
5a	O_2-sat. (sd)$<=5.8$ AND noninv-sys-BP (mean) >111.5	$3e-32$	25.07
5b	noninv-dia-BP (mean)>34.3 AND noninv-sys-BP (mean) >111.5	$4e-25$	26.30
5c	noninv-dia-BP (mean)>34.3 AND O_2-sat. (sd)$<=5.8$	$3e-60$	95.21
6	Aspartate-aminotransferase (mean)$<=2585$ AND O_2-sat. (sd)>3.1	$5e-42$	13.6
6a	Aspartate-aminotransferase (mean)$<=2580$	$5e-33$	98.97
6b	O_2-sat. (sd)>3.1	$2e-51$	14.01
7	total-los>0.52 AND heart-rate (tests) s>478 AND arterial-pH (mean)>7.25	$1e-37$	12.07
7a	heart-rate (tests) s>478 AND arterial-pH (mean)>7.25	$1e-37$	12.07
7b	total-los>0.52 AND arterial-pH (mean)>7.25	$5e-290$	90.59
7c	total-los>0.52 AND heart-rate (tests) s>478	$2e-37$	12.11
8	heart-rate (tests) >7 AND noninv-sys-BP (mean)$<=99.61$	$2e-27$	8.91
8a	heart-rate (tests) >7	1	99.4
8b	noninv-sys-BP (mean)$<=99.61$	$2e-30$	9.08

the rule at the base level. Analysis of O_2-sat. in rule 1 shows that an O_2-sat.(mean) value greater than 90 leads to an odds ratio much lower than 1, and hence decreased mortality risk. A further increase in the O_2-sat. shows a steady increase in risk showing that very

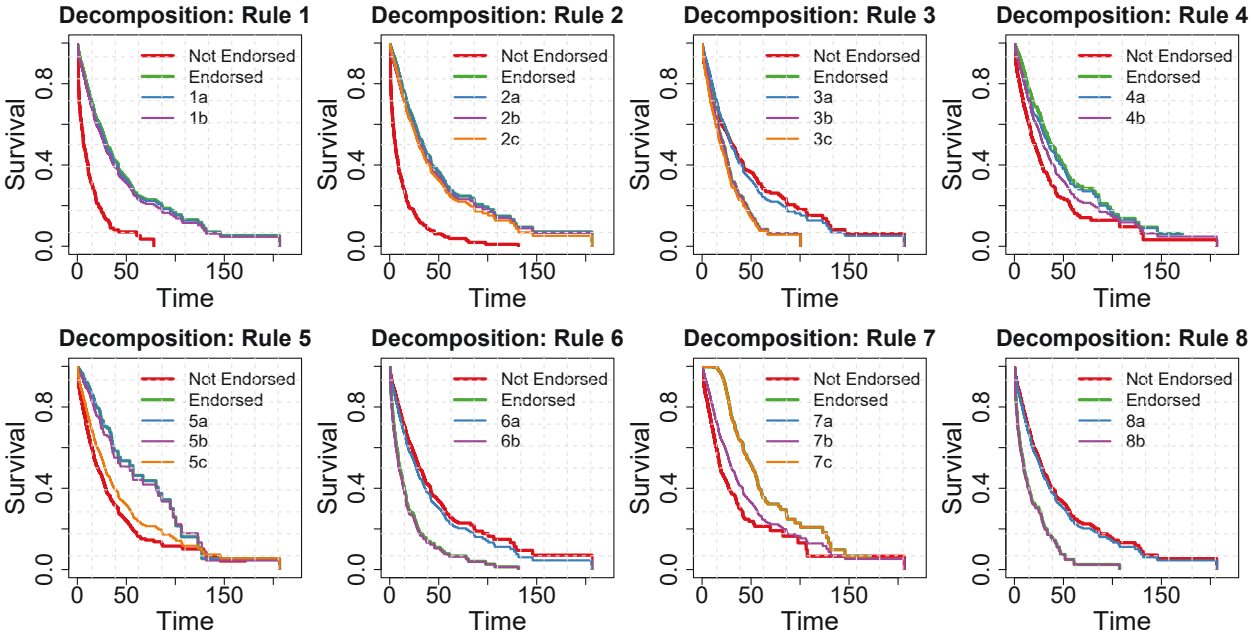


Figure 11 **Decomposition analysis curves of rules in Table 4**

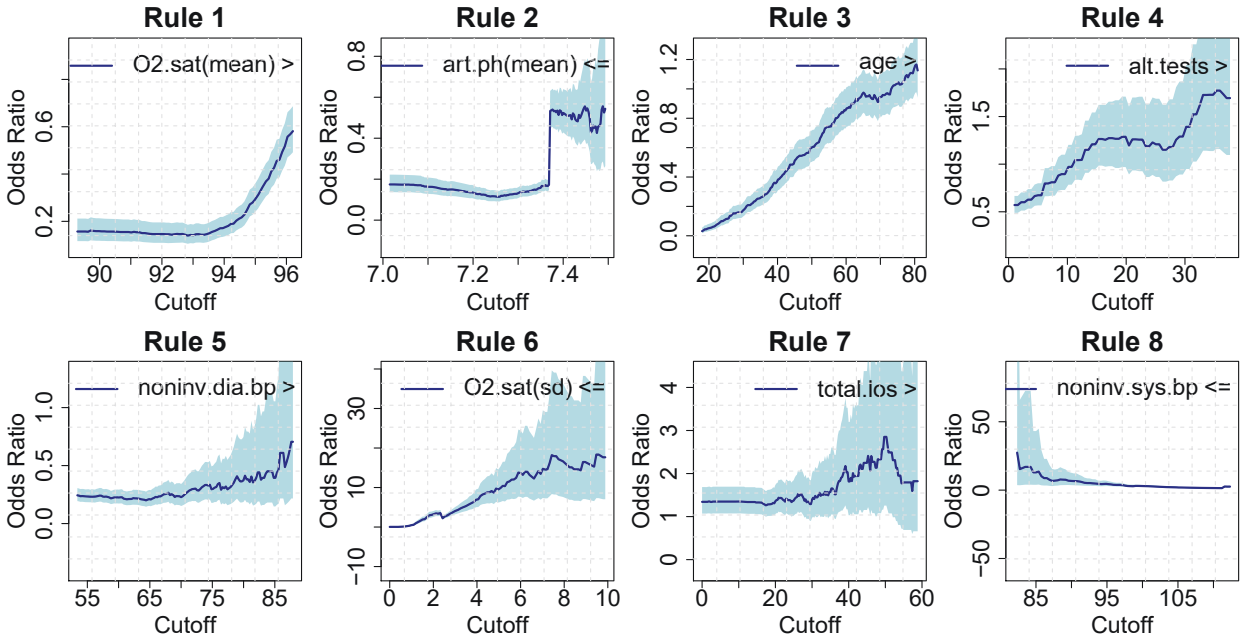


Figure 12 **Sensitivity analysis of critical factors in rules from Table 4**

high levels of oxygen saturation (above 95%) will increase mortality risk. This analysis is in line with a study by Pope et al. (2010) which reports that both abnormally high and low levels of oxygen saturation are associated with increased mortality in patients with suspected sepsis. Our sensitivity analysis is able to resolve such interactions, and predict these complex effects. In rule 2, we see that arterial-pH below 7.2 has a slightly higher

odds ratio, and therefore risk compared to when it is between 7.2 and 7.4. Any higher arterial-pH(mean) drastically increases the odds ratio implying that a high arterial-pH is a strong indicator of mortality. In rule 3, as cutoff for age increases, the odds ratio and therefore the risk of mortality of rule endorsing observations increases steadily implying that the older population is at greater risk of mortality. Advanced age has been found to be a strong predictor of mortality among sepsis patients (Yang et al. 2010, Dd et al. 1990). Rules 5 and 8 show that a high diastolic or low systolic blood pressure will increase mortality risk. These insights into the affect of blood pressure are similar to those obtained by prior research in a study conducted by Shashikumar et al. (2017) who used blood pressure and heart rate dynamics to determine risk. Meanwhile, in rule 7, we find longer length of stays are associated with higher risk until a stay of about 50 days, the large confidence interval of the odds ratio at stays which are any higher makes it hard to make a conclusion about the risk in this case.

5.1.1. Comparison with Cox Regression and Random Survival Forest. We again run experiments over 100 independently sampled subsets of the sepsis data to compare the predictive performance of the SURVFIT model with the survival random forest and the Cox model. We use 4-fold cross-validation (Hastie et al. 2009) to estimate the error rates of the 3 models being compared. This is done as follows: first we divide the dataset into 4 equal and exclusive parts. Then, one of the parts is considered the test set and the models are trained on the remaining 3 parts after which performance evaluation is done on the test set. This is done 4 times, each time considering a different part as the test set. This entire procedure is repeated 25 times for 25 different random divisions of training and the testing set to obtain the C-Index estimates for each of the different methods. The results in Fig. 13 show that, while the survival random forest and SURVFIT achieve comparable results, both methods significantly outperform the Cox model on this dataset.

To compare the differences in error rates by different methods, we use the paired Wilcoxon rank sum test on our C-Index estimates. For each pair of the methods, we perform the following test:

Null hypothesis: $\text{C-Index}_1 = \text{C-Index}_2$

Alternative hypothesis: $\text{C-Index}_1 \neq \text{C-Index}_2$

The p-Values of each of these pairwise tests are provided in Table 5. It can be seen that the performance of all 3 methods are significantly different from each other. The p-Values

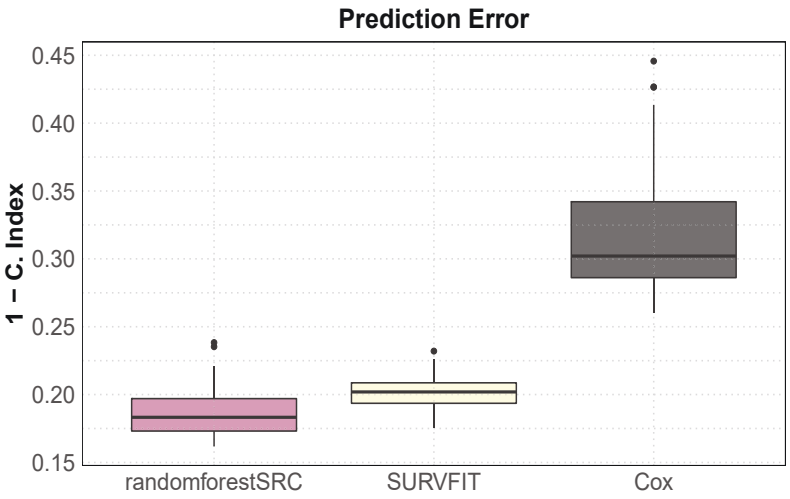


Figure 13 **Comparison of predictive performance of Random Survival Forest and SURVFIT and Cox regression on MIMIC-III Sepsis data**

of the tests show that the error rate of the Cox model is clearly higher than both randomforestSRC and SURVFIT, and the random survival forest method, randomforestSRC, achieves a lower error than SURVFIT. These results show that, on this dataset, SURVFIT yields greater interpretability than randomForestSRC at the cost of some prediction performance loss.

Table 5 **p-value of pairwise Wilcoxon rank sum test on C-index obtained by each of these methods on Sepsis data**

	randomForestSRC	Cox
SURVFIT	3.1e-05	2.98e-11
randomForestSRC		2.9e-11

6. Conclusion

Regression models dealing with survival data such as the Cox regression model are often used as confirmative tools but are limited by their inability to discover significant interaction terms from the data unless explicitly specified. This limitation is addressed by the proposed SURVFIT method which can be used to search for significant interactions among the variables. Different from existing rule learning methods, SURVFIT extracts a doubly-sparse set of rules (i.e., which are sparse both in their cardinality as well as the cardinality of the variables involved in them) for survival data. We develop the learning formulation

of SURVFIT, and further propose and evaluate fast optimization strategies. We present a rule analysis framework to analyze the extracted survival rules through statistical testing, decomposition analysis, and sensitivity analysis to draw deeper insights from them. SURVFIT could be used solely as a data analysis method that could reveal insights about the contributions and interactions of the variables. Its results could also used to augment the Cox regression as well, i.e., with higher-order interactions. Moreover, the absence of any underlying assumptions about the data makes our model quite robust. In summary, SURVFIT provides a sparse, efficient and highly interpretable tool that can be used to detect and explain the properties of predictive rules from survival data. We have also developed the R package, **SURVFIT**, to implement the rule learning algorithm and rule analysis framework presented in this paper. Future directions to SURVFIT may include development of more structured solutions, such as ones with hierarchical restrictions on the variables in the rules, as well as learning rule sets such that the rule endorsement subsets are highly unique, i.e., rules are different from each other not only in terms of the variables involved but also in terms of the observations they endorse.

7. Software and Computational Details

R package **ranger** (Wright and Ziegler 2017) was used to build survival random forest and **inTrees** (Deng 2014) was modified by us to extract an exhaustive rule list from ranger. An R implementation of FOGLASSO based on **SLEP** (Liu et al. 2009) was used to implement the first-order method. The SOCP formulation was solved using CPLEX solver. Comprehensive codes to implement solutions of both formulations, extract survival rules, and use the proposed rule-analysis framework are available in a self-contained **SURVFIT** package downloadable from `https://github.com/hamzameer/SURVFIT`.

References

Andaluz-Ojeda D, Bobillo F, Iglesias V, Almansa R, Rico L, Gandía F, Resino S, Tamayo E, de Lejarazu RO, Bermejo-Martin JF (2012) A combined score of pro- and anti-inflammatory interleukins improves mortality prediction in severe sepsis. *Cytokine* 57(3):332–336, ISSN 1043-4666, URL `http://dx.doi.org/10.1016/j.cyto.2011.12.002`.

Armijo L (1966) Minimization of functions having Lipschitz continuous first partial derivatives. *Pacific Journal of Mathematics* 16(1):1–3, ISSN 0030-8730, URL `https://projecteuclid.org/euclid.pjm/1102995080`, publisher: Pacific Journal of Mathematics.

Beck A, Teboulle M (2009) A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. *SIAM Journal on Imaging Sciences* 2(1):183–202, ISSN 1936-4954, URL http://dx.doi.org/10.1137/080716542.

Bland JM, Altman DG (2000) The odds ratio. *BMJ* 320(7247):1468, ISSN 0959-8138, URL http://dx.doi.org/10.1136/bmj.320.7247.1468.

Boyd S, Boyd SP, Vandenberghe L (2004) *Convex optimization* (Cambridge university press).

Breiman L (2001) Random Forests. *Machine Learning* 45(1):5–32, ISSN 1573-0565, URL http://dx.doi.org/10.1023/A:1010933404324.

Cendrowska J (1987) PRISM: An algorithm for inducing modular rules. *International Journal of Man-Machine Studies* 27(4):349–370, ISSN 0020-7373, URL http://dx.doi.org/10.1016/S0020-7373(87)80003-2.

Chen X, Lin Q, Kim S, Carbonell JG, Xing EP (2012) Smoothing proximal gradient method for general structured sparse regression. *The Annals of Applied Statistics* 6(2):719–752, ISSN 1932-6157, URL http://dx.doi.org/10.1214/11-AOAS514.

Cohen WW (1995) Fast Effective Rule Induction. Prieditis A, Russell S, eds., *Machine Learning Proceedings 1995*, 115–123 (San Francisco (CA): Morgan Kaufmann), ISBN 978-1-55860-377-6, URL http://dx.doi.org/10.1016/B978-1-55860-377-6.50023-2.

Cox DR (1972) Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)* 34(2):187–220, ISSN 0035-9246, URL https://www.jstor.org/stable/2985181.

Dd T, Ab G, J vdM, Jj N, Rj SvS, Lg T (1990) Age, chronic disease, sepsis, organ system failure, and mortality in a medical intensive care unit. *Critical Care Medicine* 18(5):474–479, ISSN 0090-3493, 1530-0293, URL http://dx.doi.org/10.1097/00003246-199005000-00002.

Deng H (2014) Interpreting Tree Ensembles with inTrees. *arXiv:1408.5456 [cs, stat]* URL http://arxiv.org/abs/1408.5456, arXiv: 1408.5456.

Dou J, Zhou Y, Cui Y, Chen M, Wang C, Zhang Y (2019) AST-to-Platelet Ratio Index as Potential Early-Warning Biomarker for Sepsis-Associated Liver Injury in Children: A Database Study. *Frontiers in Pediatrics* 7, ISSN 2296-2360, URL http://dx.doi.org/10.3389/fped.2019.00331, publisher: Frontiers.

Fokkema M (2017) Fitting Prediction Rule Ensembles with R Package pre. *arXiv:1707.07149 [stat]* URL http://arxiv.org/abs/1707.07149, arXiv: 1707.07149.

Ford DW, Goodwin AJ, Simpson AN, Johnson E, Nadig N, Simpson KN (2016) A Severe Sepsis Mortality Prediction Model and Score for Use With Administrative Data. *Critical care medicine* 44(2):319–327, ISSN 1530-0293, URL http://dx.doi.org/10.1097/CCM.0000000000001392.

Friedman JH, Popescu BE (2008) Predictive learning via rule ensembles. *The Annals of Applied Statistics* 2(3):916–954, ISSN 1932-6157, URL http://dx.doi.org/10.1214/07-AOAS148, arXiv: 0811.1679.

Glaab E, Bacardit J, Garibaldi JM, Krasnogor N (2012) Using Rule-Based Machine Learning for Candidate Disease Gene Prioritization and Sample Classification of Cancer Gene Expression Data. *PLOS ONE* 7(7):e39932, ISSN 1932-6203, URL http://dx.doi.org/10.1371/journal.pone.0039932.

Gotts JE, Matthay MA (2016) Sepsis: pathophysiology and clinical management. *BMJ* 353, URL http://dx.doi.org/10.1136/bmj.i1585.

Harrell J Frank E, Califf RM, Pryor DB, Lee KL, Rosati RA (1982) Evaluating the Yield of Medical Tests. *JAMA* 247(18):2543–2546, ISSN 0098-7484, URL http://dx.doi.org/10.1001/jama.1982.03320430047030.

Hastie T, Tibshirani R, Friedman J (2009) *The elements of statistical learning: data mining, inference, and prediction* (Springer Science & Business Media).

Hothorn T, Lausen B (2003) On the exact distribution of maximally selected rank statistics. *Computational Statistics & Data Analysis* 43(2):121–137, ISSN 0167-9473, URL https://mathscinet.ams.org/mathscinet-getitem?mr=1985332.

Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS (2008) Random survival forests. *The Annals of Applied Statistics* 2(3):841–860, ISSN 1932-6157, URL http://dx.doi.org/10.1214/08-AOAS169.

Jenatton R, Audibert JY, Bach F (2011) Structured Variable Selection with Sparsity-Inducing Norms. *J. Mach. Learn. Res.* 12:2777–2824, ISSN 1532-4435, URL http://dl.acm.org/citation.cfm?id=1953048.2078194.

Johnson AE, Pollard TJ, Shen L, Lehman LwH, Feng M, Ghassemi M, Moody B, Szolovits P, Anthony Celi L, Mark RG (2016) MIMIC-III, a freely accessible critical care database. *Scientific Data* 3:160035, URL https://doi.org/10.1038/sdata.2016.35.

Ke C, Jin Y, Evans H, Lober B, Qian X, Liu J, Huang S (2017) Prognostics of surgical site infections using dynamic health data. *Journal of Biomedical Informatics* 65:22–33, ISSN 1532-0464, URL http://dx.doi.org/10.1016/j.jbi.2016.10.021.

Krafft P, Steltzer H, Hiesmayr M, Klimscha W, Hammerle AF (1993) Mixed venous oxygen saturation in critically ill septic shock patients. The role of defined events. *Chest* 103(3):900–906, ISSN 0012-3692, URL http://dx.doi.org/10.1378/chest.103.3.900.

LeBlanc M, Crowley J (1992) Relative risk trees for censored survival data. *Biometrics* 48(2):411–425, ISSN 0006-341X.

Lei Yuan, Jun Liu, Jieping Ye (2013) Efficient Methods for Overlapping Group Lasso. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35(9):2104–2116, ISSN 0162-8828, 2160-9292, URL http://dx.doi.org/10.1109/TPAMI.2013.17.

Leone M, Blidi S, Antonini F, Meyssignac B, Bordon S, Garcin F, Charvet A, Blasco V, Albanèse J, Martin C (2009) Oxygen Tissue Saturation Is Lower in Nonsurvivors than in Survivors after Early

Resuscitation of Septic Shock. *Anesthesiology: The Journal of the American Society of Anesthesiologists* 111(2):366–371, ISSN 0003-3022, URL http://dx.doi.org/10.1097/ALN.0b013e3181aae72d, publisher: The American Society of Anesthesiologists.

Lin Y, Huang S, Simon GE, Liu S (2018) Data-based Decision Rules to Personalize Depression Follow-up. *Scientific Reports* 8(1):5064, ISSN 2045-2322, URL http://dx.doi.org/10.1038/s41598-018-23326-1.

Lin Y, Qian X, Krischer J, Vehik K, Lee HS, Huang S (2014) A Rule-Based Prognostic Model for Type 1 Diabetes by Identifying and Synthesizing Baseline Profile Patterns. *PLOS ONE* 9(6):e91095, ISSN 1932-6203, URL http://dx.doi.org/10.1371/journal.pone.0091095.

Liu J, Ji S, Ye J (2009) *SLEP: Sparse Learning with Efficient Projections*.

Liu V, Escobar GJ, Greene JD, Soule J, Whippy A, Angus DC, Iwashyna TJ (2014) Hospital Deaths in Patients With Sepsis From 2 Independent CohortsHospital Deaths in Patients With SepsisLetters. *JAMA* 312(1):90–92, ISSN 0098-7484, URL http://dx.doi.org/10.1001/jama.2014.5804.

Mantel N (1966) Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother Rep* 50:163–170.

Michalski RS (1980) Pattern Recognition as Rule-Guided Inductive Inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-2(4):349–361, ISSN 0162-8828, URL http://dx.doi.org/10.1109/TPAMI.1980.4767034.

Mittal S, Madigan D, Burd RS, Suchard MA (2014) High-dimensional, massive sample-size Cox proportional hazards regression for survival analysis. *Biostatistics* 15(2):207–221, ISSN 1465-4644, URL http://dx.doi.org/10.1093/biostatistics/kxt043.

Moreno RP, Metnitz B, Adler L, Hoechtl A, Bauer P, Metnitz PGH, SAPS 3 Investigators (2008) Sepsis mortality prediction based on predisposition, infection and response. *Intensive Care Medicine* 34(3):496–504, ISSN 1432-1238, URL http://dx.doi.org/10.1007/s00134-007-0943-1.

Nesseler N, Launey Y, Aninat C, Morel F, Mallédant Y, Seguin P (2012) Clinical review: The liver in sepsis. *Critical Care* 16(5):235, ISSN 1364-8535, URL http://dx.doi.org/10.1186/cc11381.

Pang H, George SL, Hui K, Tong T (2012) Gene selection using iterative feature elimination random forests for survival outcomes. *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM* 9(5):1422–1431, ISSN 1545-5963, URL http://dx.doi.org/10.1109/TCBB.2012.63.

Patil BM, Joshi RC, Toshniwal D (2010) Association Rule for Classification of Type-2 Diabetic Patients. *2010 Second International Conference on Machine Learning and Computing*, 330–334, URL http://dx.doi.org/10.1109/ICMLC.2010.67.

Paulon G, De Iorio M, Guglielmi A, Ieva F (2020) Joint modeling of recurrent events and survival: a Bayesian non-parametric approach. *Biostatistics* URL http://dx.doi.org/10.1093/biostatistics/kxy026.

Pope JV, Jones AE, Gaieski DF, Arnold RC, Trzeciak S, Shapiro NI (2010) Multicenter Study of Central Venous Oxygen Saturation (ScvO2) as a Predictor of Mortality in Patients With Sepsis.

Annals of Emergency Medicine 55(1):40–46.e1, ISSN 0196-0644, URL http://dx.doi.org/10.1016/j.annemergmed.2009.08.014.

Quinlan JR (1990) Learning logical definitions from relations. *Machine Learning* 5(3):239–266, ISSN 1573-0565, URL http://dx.doi.org/10.1007/BF00117105.

Segal MR (1988) Regression Trees for Censored Data. *Biometrics* 44(1):35–47, ISSN 0006-341X, URL http://dx.doi.org/10.2307/2531894.

Shashikumar SP, Stanley MD, Sadiq I, Li Q, Holder A, Clifford GD, Nemati S (2017) Early sepsis detection in critical care patients using multiscale blood pressure and heart rate dynamics. *Journal of Electrocardiology* 50(6):739–743, ISSN 0022-0736, URL http://dx.doi.org/10.1016/j.jelectrocard.2017.08.013.

Simon N, Friedman J, Hastie T, Tibshirani R (2013a) A Sparse-Group Lasso. *Journal of Computational and Graphical Statistics* 22(2):231–245, ISSN 1061-8600, URL http://dx.doi.org/10.1080/10618600.2012.681250.

Simon N, Friedman J, Hastie T, Tibshirani R (2013b) A Sparse-Group Lasso. *Journal of Computational and Graphical Statistics* 22(2):231–245, ISSN 1061-8600, URL http://dx.doi.org/10.1080/10618600.2012.681250.

Sweeney TE, Perumal TM, Henao R, Nichols M, Howrylak JA, Choi AM, Bermejo-Martin JF, Almansa R, Tamayo E, Davenport EE, Burnham KL, Hinds CJ, Knight JC, Woods CW, Kingsmore SF, Ginsburg GS, Wong HR, Parnell GP, Tang B, Moldawer LL, Moore FE, Omberg L, Khatri P, Tsalik EL, Mangravite LM, Langley RJ (2018) A community approach to mortality prediction in sepsis via gene expression analysis. *Nature Communications* 9(1):694, ISSN 2041-1723, URL http://dx.doi.org/10.1038/s41467-018-03078-2.

Tarkhan A, Simon N (2020) Bigsurvsgd: Big survival data analysis via stochastic gradient descent.

Tibshirani R (1996) Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 58(1):267–288, ISSN 0035-9246, URL https://www.jstor.org/stable/2346178.

Tibshirani R (1997) The Lasso Method for Variable Selection in the Cox Model. *Statistics in Medicine* 16(4):385–395, ISSN 1097-0258, URL http://dx.doi.org/10.1002/(SICI)1097-0258(19970228)16:4<385::AID-SIM380>3.0.CO;2-3.

Wang P, Li Y, Reddy CK (2019) Machine learning for survival analysis: A survey. *ACM Computing Surveys (CSUR)* 51(6):1–36.

Wright MN, Ziegler A (2017) ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software* 77(1):1–17, ISSN 1548-7660, URL http://dx.doi.org/10.18637/jss.v077.i01.

Wróbel Gudyś A, Sikora M (2017) Learning rule sets from survival data. *BMC Bioinformatics* 18, ISSN 1471-2105, URL http://dx.doi.org/10.1186/s12859-017-1693-x.

Yang Y, Yang KS, Hsann YM, Lim V, Ong BC (2010) The effect of comorbidity and age on hospital mortality and length of stay in patients with sepsis. *Journal of Critical Care* 25(3):398–405, ISSN 0883-9441, URL http://dx.doi.org/10.1016/j.jcrc.2009.09.001.

Yuan M, Lin Y (2006) Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68(1):49–67, ISSN 1369-7412, 1467-9868, URL http://dx.doi.org/10.1111/j.1467-9868.2005.00532.x.

Zagory JA, Dossa A, Golden J, Jensen AR, Goodhue CJ, Upperman JS, Gayer CP (2017) Re-evaluation of liver transaminase cutoff for CT after pediatric blunt abdominal trauma. *Pediatric Surgery International* 33(3):311–316, ISSN 1437-9813, URL http://dx.doi.org/10.1007/s00383-016-4026-7.