STRUCTURED RULE DISCOVERY FROM HETEROGENEOUS LONGITUDINAL DATA FOR COMPLEX DISEASE

A Thesis

by

ZHOU WANG

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirement for the degree of

MASTER OF SCIENCE

Chair of Committee,	Peng Li
Co-Chair of Committee,	Xiaoning Qian
Committee Members,	I-Hong Hou
	Xia Hu
Head of Department,	Miroslav Begovic

December 2016

Major Subject: Computer Engineering

ABSTRACT

Many complex diseases manifest heterogeneous degenerative disease progression processes that impose enormous challenges for accurate disease prognosis and effective intervention. The emerging "big" data collected from the population with predisposed disease risk brings us motivation to translate it into accurate prognosis and effective risk monitoring. We propose a structured sparse rule discovery method to identify risk-predictive patterns from heterogenous longitudinal. By extending the existing RuleFit framework, we have developed an analysis pipeline to derive risk-predictive patterns from complex data. The results in a de-identified Type 1 Diabetes dataset have shown promising predictive performances.

This thesis is dedicated to my mother and father, who have supported me with love.

ACKNOWLEDGMENTS

Dr. Xiaoning Qian has been a great supervisor. His sage advice, insightful criticisms, and patient encouragement aided me with the whole process of this thesis in every way. I would like to thank Dr. Peng Li, Dr. I-Hong Hou, and Dr. Xia Hu whose support of this project was greatly needed and deeply appreciated. I am also grateful for the help and advice from Dr. Shuai Huang at the University of Washington. The help from Drs. Kendra Vehik and Kristian Lynch is truly appreciated for the data from The Environmental Determinants of Diabetes in the Young (TEDDY) study.

NOMENCLATURE

AD	Alzheimer’s Disease
PD	Parkinson Disease
T1D	Type 1 Diabetes
RF	Random Forest
LASSO	Least Absolute Shrinkage and Selection Operator
OGL	Overlapped Group LASSO
OGAPS	Office and Graduate and Professional Studies at Texas A&M University
TAMU	Texas A&M University

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

1 INTRODUCTION

1.1 Background

Many degenerative complex diseases, such as autoimmune disorders (e.g., Type 1 Diabetes, Celiac Disease, Rheumatoid Arthritis, Multiple Sclerosis), chronic disease (e.g., Asthma, Cardiac Diseases), and neurodegenerative diseases (e.g., Alzheimer's Disease (AD), Parkinson Disease (PD), Amyotrophic Lateral Sclerosis), usually exhibit heterogenous phenotypes. The exact causes of those complex diseases are still in debate. However, the increasing evidences in the literature [1, 2], indicate that the evolution to disease clinical onsets can be affected by various exogenous factors, such as different early-life environment exposures including dietary as well as infection exposures. It is possible to intervene to delay [3], or even avoid the disease clinical onset if we can better model, understand, and predict the disease progression processes by identifying potential risk-predictive patterns from high-dimensional measurements of candidate genetic, biomolecular, clinical, as well as environmental risk factors.

Discovering risk-predictive factors residing in massive biomedical data that have been accumulated in existing clinical trials, is potentially helpful but also imposes enormous data analytic challenges. Take Type 1 Diabetes (T1D) as an example, several studies have been conducted to collect large scale data for risk factor identification for T1D data. Translating such big and complex data to accurate and reproducible understanding complex diseases requires appropriate data analytic methods. One frequently applied method is to analyze risk variables over observable outcomes with logistic regression [4]. However, common logistic regression models generally examine average risk effects over whole group without considering complex structured patterns, which might hide in the subgroups. Therefore, a rule-based analysis approach has recently been proposed to identify baseline profile patterns, and synthesize these pattern for risk prediction [5]. Such rule-based analyses can help model potential nonlinear interactive effects among candidate factors for better risk-predictive factor identification. However, besides baseline profile factors, temporal changes from the collected longitudinal data that can also potentially be risk-predictive have not been carefully studied.

1.2 Problem Statement

The problem to be studied in this thesis is to develop a structured rule-based method to derive risk-predictive patterns from data featured with temporal measurements. The problem fits within the ongoing data-driven discovery to model, understand, predict, and treat complex diseases with the help of mixed types of data collected with heterogeneous binary, categorical, quantitative measurements at the baseline as well as across disease progression time.

There are a number of essential challenges in this study. First of all, we must integrate mixed types of measurements into a unified statistical model to identify risk-predictive patterns in a more systematic way. Beyond that, with longitudinal data, it is critical to appropriately model the dependency structures among temporal measurements for the corresponding factors that may change over time. Finally, with high-dimensional measurements, it is necessary to select non-redundant risk-predictive patterns to avoid potential overfitting problems in risk factor identification.

To overcome these challenges, we integrate rule-based analysis [6] and Overlapped Group LASSO (OGL) [7] in this study.

Rule-based analysis brings some special advantages, in comparison with earlier studies using logistic regression. First, deriving rules can be considered as a normalization step, which can convert diverse data types into consistent binary indicators in a supervised fashion. This naturally handles the problem of having mixed types of data and helps integrate binary, categorical, and numeric variables or features by mapping them to binary outputs based on whether their values fall into interesting ranges with respect to the outcome of interest. Second, the derivation rules are dependent on derived value ranges by setting thresholds considering the disease risk outcome. This makes it possible to model nonlinearity that is relevant to disease risk. Finally, the derived rules often involve two or more features. It incorporates interactions among features for risk prediction.

Applying Overlapped Group LASSO with rule-based methods helps to facilitate the flexibility to model the group dependency among rules and candidate factors. When longitudinal data is to analyze with potential temporal dependecy, such consideration of stuctured dependency may help derive more robust and accurate risk-predictive patterns while avoiding overfitting to ensure that the identified rules are non-redundant and significant.

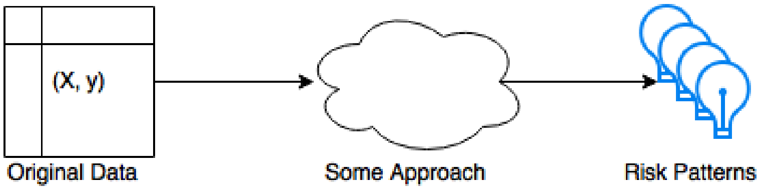


Figure 1.1: We want an approach to translate disease data into predictive patterns.

The purposes of this study are to 1) design a workflow(see Figure 1.1) to generate rules and appropriately select top rules as risk-predictive patterns for disease prognosis and risk monitoring to better understand and intervention of complex degenerative diseases; 2) evaluate the rule-based methods when deriving complex disease progression patterns from big, complex, and heterogenous data.

2 LITERATURE REVIEW

Biomarker and risk factor discovery has been studied extensively in the literature. Different statistical and computational methods have been proposed to analyze collected measurements to derive potential factors that may help understand and predict disease progression and outcomes of interest.

2.1 Biomarker Discovery and Disease Prognosis

Logistic regression model has been the standard biostatistics method applied to study the association between disease outcome and candidate risk factors. As early as 1983, T.A. Welborn, *et al.* conducted a research [4] to identify risk factors for clinical macro-vascular diseases among diabetic subjects by applying a logistic regression method to determine the strength of association of possible risk variables over the outcome, clinical onset of macro-vascular diseases. In such association analyses, the logistic regression was performed for each variable, and the association was evaluated by the coefficient p-values, by which a subset of variables can be selected as "biomarkers" or "risk factors" considering the deviance contribution derived from the regression.

Logistic regression also has been implemented for disease prognosis. Another study in 2002 [8], by Bahman P. Tabaei, *et al.*, proposed an empirical predictive equation for diabetes screening by multiple logistic regression analysis. They collected data from subjects with no history of diabetes, assessed baseline characteristics, such as age, gender, height, weight, postprandial time, and estimated the likelihood of previously undiagnosed diabetes based on these candidate risk variables. The researchers finally developed a screening method for diabetes based on a predictive equation.

The past attempts to solve biomarker discovery and disease prognosis problems by conducting logistic regression analysis have been mostly hypothesis-driven, by analyzing a limited number of carefully selected observable risk factors. For example, Welborn's research validated that age is the major risk factor for macro-vascular disease among both Type 1 and Type 2 diabetic subjects, and plasma creatinine levels and plasma glucose levels are also significant risk predictors for Type 2 Diabetes subjects. However, many complex diseases are conjectured to be caused by both genetic and environmental risk factor exposures, as well as their interactions. Hypothesis-driven analyses may not reveal the disease progression mechanisms without accurate and systematic identification of all influencing factors. For example, in Welborn's study, they failed to validate some well-recognized risk factors for vascular diseases, such as cigarette smoking. Although they claimed that this discrepancy may be caused by the crudity of simple smoking questionnaire, ignoring interactive effects in traditional association analyses by logistic regression in their research may be one of the reasons that they failed to demonstrate the smoking effects on macro-vascular complications for diabetic subjects.

In order to integrating interaction in addition to individual effects for risk factor identification and disease prognosis, one way is to perform new logistic regressions on subgroups of the variables including their potential interaction terms. However, with the current trend of data-driven research to collect large-scale and high-dimensional observable measurements in biomedical studies, we face enormous data analytic challenges to do that. First, adding interaction terms can increase the model complexity and computational complexity in an exponential fashion. With often a limited number of samples in biomedical studies, it may lead to many false discoveries in addition to computational difficulty. When adding interaction terms with the help of experts or existing hypotheses, it may incur the cost from human labor and potential bias towards the analysis results. Second, with the availability of different types of measurements, it requires appropriate analytic methods to integrate all different types of data in one unified analysis framework to obtain reproducible, interpretable, and meaningful results. Finally, when we have longitudinal measurements from prospective clinical trials, how to model the temporal dependency imposes another potential challenge.

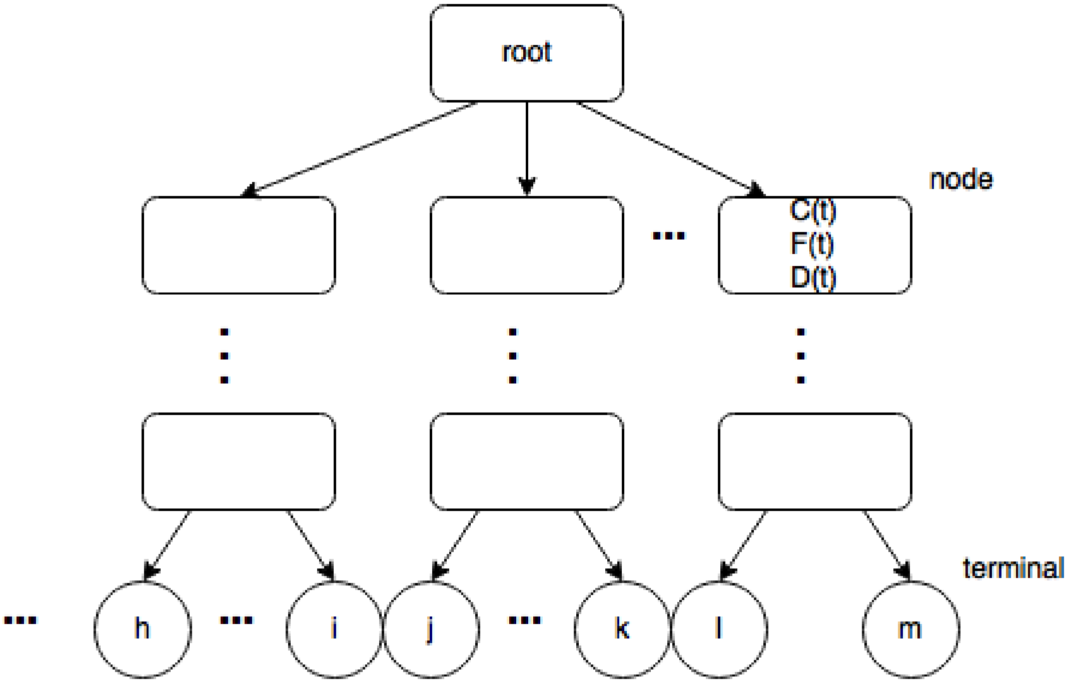
To overcome these analytic challenges, we adopt rule-based methods for risk factor identification when analyzing complex and heterogeneous data from biomedical studies. Rule-based methods [9] are based on decision trees. In the following, we give the literature review on decision tree, random forest, and rule-based analytic methods and their applications in biomedical research.

2.2 Decision Tree

The decision tree classifiers are widely and successfully applied in diverse areas, e.g., biomedical engineering, financial analysis, system controls, software engineering. As stressed in S.R. Safavian's survey paper [10], the most crucial property of a decision tree model is its capability to simplify complex decision-making process by a representation of sequences of comprehensible decision rules. A decision tree is a tree-like directed graph model, where each tree node contains the corresponding trained decision-making rule. The basic decision-making scenario is that when a candidate sample comes in, the final decision is made by starting from the root of the trained decision tree, going down to next one until the leaf node based on the hierarchical decision rules along the path.

A general decision Tree T for decision-tree classifiers can be illustrated in Figure 2.1, where each tree node t contains three key factors: 1) accessible class subset from t, 2) feature subset, and 3) decision rule applied in t. Decisions would be made when candidate testing samples reach leaf nodes at the terminal level, and are assigned with the corresponding class labels.

Since it has been proved [11] that constructing the optimal binary decision tree is NP-Complete problem, it is computationally prohibitive when the data dimension grows.



$C(t)$ - accessible class subset from node t
$F(t)$ - feature subset applied at node t
$D(t)$ - decision rule applied at node t

Figure 2.1: Illustration of a general decision tree.

Heuristics based on some common criteria are adopted for building the decision tree classifiers based on training samples, including minimum error rate and minimum number of tree nodes, etc. For example, in bottom-up training, training set samples are iteratively grouped as tree nodes with member samples having small distances, measured by selected metrics, until we form the root node of the tree. For top-down methods, the iterative procedure can be decomposed into three steps: 1) selecting optimal splitting rules based on the criterion of the choice, 2) assigning class label for the corresponding tree nodes, and 3) determining terminating criterion.

For decision tree classifier design on finding node splitting rules, there are multiple measures of evaluating the quality of splitting [12]. Two of the most popular indices include (1) Gini Impurity, which is implemented in CART (Classification And Regression Tree), and (2) Information Gain, which is implemented in ID3 (Iterative Dichotomiser 3) and its extension C4.5.

Let T denote an accessible subset of class in the form of $(\mathbf{x}, y) = (x_1, x_2, \ldots, x_k, y)$, where x_a is the value of the a-th attribute or factor, $y \in \{1, 2, \ldots, J\}$ is the class label.

Gini impurity is defined as the probability of incorrectly assigning class labels based on the class distribution for randomly selected samples in a given node. In each split decision making, the algorithm in CART seeks for a subsequent condition where Gini impurity is minimized. It takes this form:

$$I_G(T) = 1 - \sum_{i=1}^{J} P(y=i)^2 = \sum_{i \neq k} P(y=i)P(y=k), \tag{2.1}$$

where $P(y=i)$ indicates probability of label of random item in set is i.

Information gain is the decrement amount of information entropy H from a state to another. Algorithms applying this metric seek for the optimal information gain in each split.

The entropy in the splitting node is

$$H(T) = -\sum_{i=1}^{J} P(y=i) \log P(y=i) \tag{2.2}$$

The information gain by setting splitting according to $x_a = v$ is

$$IG(T, a, v) = H(T) - \sum_{v \in vals(a)} \frac{|\mathbf{x} \in T \mid x_a = v|}{|T|} \cdot H(\mathbf{x} \in T \mid x_a = v) \tag{2.3}$$

Another interesting related issue in the design phase is the missing value problem. For training samples with missing feature, a simple approach is to throw them away, but it is not applicable with testing samples. The common solution in decision trees is called surrogate split. The basic idea is that the best split at current node S^* is known, then we can find another splitting criterion $\hat{S}$ without considerring missing features where similarity of two splits $\lambda(S^*, \hat{S})$ are maximized, where the surrogate splits $\hat{S}$ divide the training dataset, and leads to a similar result as divided by optimal split $\lambda(S^*, \hat{S})$. Then we may use $\hat{S}$ as to determine whether the incoming test samples should traverse to the left or right child.

In general, there are some positive characteristics for decision-tree-like models: 1) the capability of handling numerical and categorical attributes together, 2) less requirements for data preprocessing, 3) a “white-box” model, where decision rules are explicit and can be extracted as interpretable patterns, 4) good performance for large datasets because of efficient heuristic search strategies.

Decision trees have been implemented to to discover predictive patterns for complex disease prognosis [13, 14] in the past to analyze observable biomedical measurements and derive rules with decision tree models to predict the disease outcome. These approaches have been reported to achieve high prediction accuracy with flexible and interpretable structured patterns. It can be a better option, compared to logistic regression based methods when we want to underline variable associations, and model the structure patterns as rules.

2.3 Random Forest

To further improve the performances of decision tree methods, there have been two common ensemble learning approaches to construct prediction models with multiple decision trees: boosting [15] and bagging [16]. In boosting, the algorithm repeatedly constructs decision trees on randomly sampled subsets of training data, and tends to assign heavier weights to those training samples misclassified by the existing trees when generating new trees, to construct a weighted ensemble for prediction. While constructing multiple decision trees in bagging, each tree is generated independently with bootstrap sampling training data for multiple times. After that, the outcome of a sample x' can be predicted as either the majority of the decision tree votes or by this following the ensemble rule:

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b(x') \tag{2.4}$$

Here, f_b denotes the decision tree constructed on the b-th subset of training data, $b =$

$1,2,\ldots,B$. By the bootstrapping procedure, and constructing numerous trees with similar but not the exactly same training data, it significantly reduces variances and sensitivity to noisy samples, and therefore increases the robustness of the ensemble model.

Random Forests (RF) algorithm is proposed by Breiman [17] based on the bagging approach. Besides bootstrapping the training set over samples, RF increases its randomness by splitting nodes based on random selection or linear combination of features, rather than on all possible features in analyses. By introducing both sample and feature randomness, RF does not only reduce overfitting because of the law of large numbers, but also produces good prediction accuracy for classification purposes.

Random forest analyses have also been adopted to perform complex disease prediction [18], in which researchers applied RF to analyze National Inpatient Sample data, divided the data into balanced subgroups to construct decision trees, then predicted eight disease categories. And they claimed that their method performs slightly better than other popular methods such as support vector machines, bagging and boosting methods with linear classifiers.

Random forest can be a suitable approach to perform our task to discover risky patterns in complex disease data, since it does not only enable highly accurate prediction, but also provides interpretable structured rules, which can be good representations for risk-predictive patterns. For the next steps for pattern discovering, if we are able to find a good way to select the rules of high significance and high confidence from the entire rule set, we can conclude risk-predictive patterns with exactly that subset of rules.

2.4 Rule-Based Learning Ensembles

In addition to tree ensemble learning, there is also predictive learning approach via rule ensembles [6] based on RF. Rules are defined as directed paths on decision trees. It can be obtained by traversing decision trees generated with random forests. Instead of simply averaging the predictions of ensemble trees as in random forest prediction, it represents the predictive model as

$$F(\mathbf{x}) = \hat{a}_0 + \sum_{k=1}^{K} \hat{a}_k r_k(\mathbf{x}), \tag{2.5}$$

where $r_k(\mathbf{x})$ denotes the output of the corresponding sample $\mathbf{x}$ applied to the k-th rule, $k \in \{1, 2, \ldots, K\}$, and $\hat{a}_k$ is the estimated weight for the k-th rule.

Learning the coefficients a_k is achieved by a regularized linear regression over N samples in training data, with both prediction loss function and l_1-norm of the coefficient weight vector $\{\hat{a}_k\}_0^K$ considered.:

$$\{\hat{a}_k\}_0^K = argmin_{\{\hat{a}_k\}_0^K} \sum_{i=1}^{N} L(y_i, a_0 + \sum_{k=1}^{K} a_k r_k(\mathbf{x}_i)) + \lambda \cdot \sum_{k=1}^{K} |a_k|, \tag{2.6}$$

where the first term $\sum_{i=1}^{N} L(y_i, F(\mathbf{x}))$ measures the empirical prediction risk, which often takes the least-square form for regression and the logistic loss function for classification, and the second term denotes the sparsity regularize term to enable the selection of significantly contributing rules when predicting outcome responses. Such a formulation is known as Least Absolute Shrinkage and Selection Operator (LASSO) [19], which we will discuss in the next section. The λ balances between the risk and regularization terms. By taking a larger value of λ, the optimization generally yields more sparse coefficient vector $\{\hat{a}_k\}_0^K$. Thus, a subset of weighted rules finally ensembles the prediction.

Somethings else interesting about this paper [6] is that they also proposed measurement criteria for the importance of individual rule predictor, denoted as I_k. For each rule, its global importance can be measured as

$$I_k = |\hat{a}_k| \cdot \sqrt{s_k(1 - s_k)}, \tag{2.7}$$

where s_k is the rule support, characterizing how much percentage of the training samples satisfying the corresponding rule r_k. And for its local measure for any testing data sample $\mathbf{x}$, it can be computed as

$$I_k(\mathbf{x}) = |\hat{a}_k| \cdot |r_k(\mathbf{x}) - s_k| \tag{2.8}$$

Although the importance measurement does not necessarily reveal the quality of individual rules without referring to other dependencies, the information importance I_k is a good estimation for a given rule's capability of distinguishing different samples, which might contain potential information to narrow the range of patterns for further validation.

RuleFit [9] package implements a predictive learning method via rule ensembles [6]. It performs rule generation with RF by considering each decision tree node as an individual rule, and perform rule selection with the LASSO formulation [19]. With the LASSO regularization term, we can control the size of the selected rule subset to avoid overfitting and identify significant rules as disease predictive patterns, still minimizing the overall prediction risk as in logistic regression.

We have recently implemented the rule-discovery algorithm, RuleFit, to discover complex disease risk patterns. For instance, Lin, *et al.* conducted a research [5] with RuleFit on Diabetes Prevention Trial-Type 1 data to explore whether some selected basic biomedical markers can reveal significant information to predict the risk of diabetes clinical onset. They identified risk-predictive rules by generating rules with Random Forest, and selecting rules with LASSO. And the derived risk-predictive rules are consistent with results in the existing literature on T1D. In addition, Haghighi, *et al.* compared RuleFit to logistic regression analysis in exploring risk factors in dropping out from a T1D prospective study [20]. They found that both methods yield similar risk factors, but RuleFit is more beneficial because of its capability to detect multi-risk-factor interactions.

2.5 LASSO

LASSO [19] is a sparse linear regression model featured with its capability to select a subset of significant variables as potential risk factors from a large collection of candidate variables. The LASSO estimator $\hat{\beta}_\lambda$ was proposed originally for linear regression problems:

$$argmin_{\beta,\mu}||Y-\mu\cdot\mathbf{1}-\mathbf{X}\beta||_2^2+\lambda\sum_{j=1}^{p}||\beta_j||_1, \tag{2.9}$$

where $Y\in\mathbf{R}^n$ denotes the outcome response, $\mathbf{X}$ denotes a $n\times p$ sample-feature matrix, with the penalty parameter λ, μ is the intersept.

LASSO with l_1-norm regularization enables penalized feature subset selection when analyzing high-dimensional measurements, and ensures prediction stability as in ridge regression, which takes the $||\beta_j||_2^2$ regularization term instead of $||\beta_j||_1$. Due to the geometric nature of the l_1 norm, more components in β shrink to exactly zero as λ is set larger. This property is especially helpful during feature selection. As discussed earlier, this formulation is applied in the RuleFit workflow, which is an implementation of rule-based learning ensembles [6].

There are a few LASSO variant models proposed as its generalization. And they can be especially powerful for addressing specific problems, especially when dependency structures among predictors need to be taken care. The typical way of achieving that, is to add or modify the regularization terms, so that predictor associations can be modeled and considered during optimization. For example, the elastic net [21] re-designs the optimization function by adding a ridge regression regularization l_2^2 term. It becomes

$$argmin_{\beta,\mu}||Y-\mu\cdot\mathbf{1}-\mathbf{X}\beta||_2^2+\lambda_1||\beta||_1+\lambda_2||\beta||_2^2 \tag{2.10}$$

It addresses one of the limitations of LASSO, where the number of features is greater than the sample size, only the most significant features can be selected from the highly associated set. It combines the benefit of LASSO and ridge penalty, so that contribution feature subset can be better selected, in which correlated features are more likely to have similar regression coefficients.

Another LASSO variant is known as fused LASSO [22], also proposed by Tibshirani, which is designed to address the dependency problems when analyzing spatial and temporal covariates. It takes this form

$$argmin_{\beta,\mu}||Y - \mu \cdot \mathbf{1} - \mathbf{X}\beta||_2^2 + \lambda_1 \sum_{j=1}^{p} ||\beta_j||_1 + \lambda_2 \sum_{j=2}^{p} ||\beta_j - \beta_{j-1}||_1 \tag{2.11}$$

The second term is introduced to enforce the similarity of fitting coefficients between dependent covariates, thus selected features tend to distribute in multiple continuous small regions. This can be useful to model underlying patterns within one-dimentional sequential covariates. In addition, there are other generalizations of fused LASSO, including cluster LASSO by replacing the second penalty term with $\lambda_2 \sum_{i<j}^{p} |\beta_i - \beta_j|$.

Group LASSO is introduced by Yuan and Lin in 2006 [23]. It meets the purpose of applying prior knowledge on feature groups and selecting or not selecting groups together. The resulting optimization problem is:

$$argmin_{\beta,\mu}||Y - \mu \cdot \mathbf{1} - \sum_{j=1}^{J} \mathbf{X_j}\boldsymbol{\beta_j}||_2^2 + \lambda \sum_{j=1}^{J} \sqrt{p_j}||\beta_j||_2, \tag{2.12}$$

where X_j indicates feature measurements assigned to the j-th group, and $\sqrt{p_l}$ is determined by the group size.

A sparse group LASSO [7] is a further extension from group LASSO. It also produces sparse models within a group, in addition to enforcing group association. It is achieved by involving both group LASSO penalty and basic lasso penalty, and takes this form

$$argmin_{\beta,\mu}||Y - \mu \cdot \mathbf{1} - \sum_{j=1}^{J} \mathbf{X_j}\boldsymbol{\beta_j}||_2^2 + \lambda_2 \sum_{j=1}^{J} \sqrt{p_j}||\beta_j||_2 + \lambda_1 ||\beta||_1, \tag{2.13}$$

where $j \in \{1, 2, \ldots, J\}$.

The group LASSO with overlap is also proposed, where features are potentially assigned into multiple overlapping groups. It allows greater flexibility than typical group LASSO to predefine various complex structured interactions. Since an efficient algorithm [24] has been developed, overlapping group LASSO becomes an powerful approach when analyzing high-dimensional measurements of interacting factors in practice, which is often the case when we analyze big and complex biomedical data.

3 METHODS

In this chapter, we describe our structured rule-based approach for risk-predictive pattern discovery by analyzing heterogeneous longitudinal complex data. It extends the existing RuleFit [6] approach to specifically address the challenge when analyzing longitudinal measurements, where some measurements of candidate risk factors are collected along time and hence, the corresponding features manifest as two-dimensional time-variant measurements. The essential idea of structured sparse rule discovery is to modify the RuleFit rule pruning procedure to replace the step with the LASSO formulation, with the overlapped group LASSO (2.13), so that we can enforce temporal dependency structures by group regularization terms.

3.1 Problem Statement

In biomedical applications, it is often valuable but difficult to fully understand complex disease progression patterns. With the development of efficient high-throughput profiling techniques, powerful computational resources, and efficient machine learning algorithms, researchers hope to analyze large-scale data collected from biomolecular, clinic, and daily-life observations to seek for accurate and reproducible identification of potential risk factors that may affect and trigger disease development so that more accurate disease prognosis and effective therapeutics can be developed based on them. We propose here a structured sparse rule discovery method to identify risk-predictive patterns as composite biomarkers by analyzing large-scale heterogeneous longitudinal data. The proposed method takes advantages of the existing advanced rule-based and structured sparse regularization methods to specifically address the critical challenges in risk-predictive pattern discovery: (1) It naturally handles the data challenges, including diverse data types and

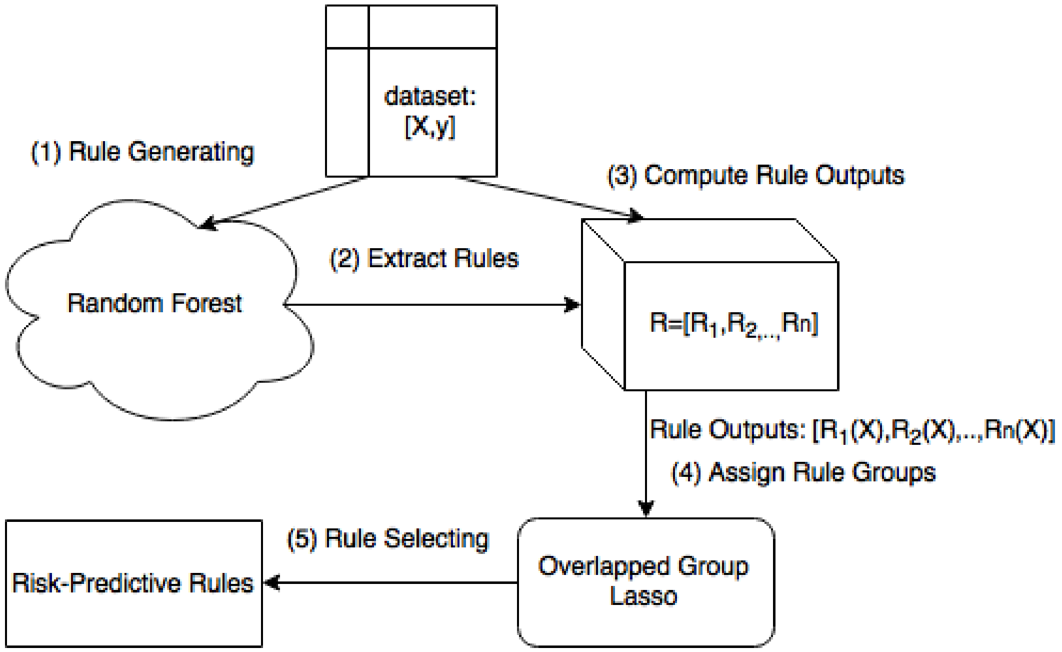


Figure 3.1: Proposed Pipeline

potential missing values in collected measurements; (2) It incorporates the potential interactive effects among different candidate factors by employing rule-based methods; (3) It aims to select non-redundant and significant rules as risk patterns that alleviates the potential overfitting problem and addresses the interpretability issues by sparse rule pruning methods; (4) Last but not least, incorporating overlap group LASSO regularization helps to take care of potential dependency structures among candidate factors and temporal dependency when analyzing longitudinal data.

The proposed structured spare rule discovery analysis pipeline is illustrated in figure 3.1.

3.2 Data Representation

As typical in feature selection and predictive learning problems, we denote the outcome label by $\mathbf{y}$, which is evaluated by a candidate feature set $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_f]$, where

f is the number of candidate factors that we can collect the corresponding measurements. In biomedical studies, these candidate factors may interact with each other and influence the disease progression process. In addition, some factors may change over time when disease develops. The corresponding measurements manifest as longitudinal data for these dynamic features. For clear representations, we assume that the longitudinal measurements are aligned with time points, thus we can have a subset in $\mathbf{X}$ in addition to static (or baseline) features: $\hat{\mathbf{X}} = [\mathbf{X}_{(1,1)}, \ldots, \mathbf{X}_{(1,q)}, \mathbf{X}_{(2,1)}, \ldots, \mathbf{X}_{(2,q)}, \ldots, \mathbf{X}_{(p,1)}, \ldots \mathbf{X}_{(p,q)}]$, where there are p dynamic features, observed at q sequential discrete time points. Hence, with both baseline and longitudinal measurements, we can represent all of them by a design matrix $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_f, \mathbf{X}_{(1,1)}, \ldots, \mathbf{X}_{(1,q)}, \ldots, \mathbf{X}_{(p,1)}, \ldots \mathbf{X}_{(p,q)}]$.

In this work, we want to derive rules as a representation of risky-predictive patterns by extending RuleFit to the integrated framework in figure 3.1 to seek for better performance on analyzing such heterogenous longitudinal data.

3.3 Structured Sparse Rule Discovery

As illustrated in figure 3.1, our rule-based approach consists of two major steps, rule generation and rule pruning.

3.3.1 Rule Generation

A rule is defined as an indicator function R, which maps the corresponding values that a feature vector $\mathbf{x}$ can take onto a binary output by checking whether a specified condition of the involved features is satisfied, often the observed measurements being within the specified range. A general rule function take one or more features and these features can be of different data types with binary, categorical or numerical values. We can write a rule as $R(\mathbf{x}) = I(\mathbf{x}_1 \in S_1) \cdot \cdots \cdot I(\mathbf{x}_f \in S_f)$, where $I(\cdot)$ is an indicator function, S_f denotes the desired value range of the f-th feature, no matter whether it is based on discrete categorical values or continuous real values.

In the first rule generation stage, we take advantage of the existing fast random forest algorithm [17] to generate rules by fitting decision tree models with the training data. Random forest serves as an extended decision tree model. It ensembles a large number of decision trees estimated from the training data through the bootstrap aggregating algorithm featured with random sampling from both subsets of samples and features. The random forest procedure helps to construct decision tree ensembles with subpopulations fully searched, and potential interactions among features are fully explored and evaluated to pick the rules with interacting features with a high likelihood as candidate risk factors when forming node splitting criteria. In this way, the full set of rules resulting from random forest can be considered informative and candidate risk-predictive patterns to describe the characteristics of the training dataset.

In each random decision tree from the derived random forest, each non-root tree node has a rule representation and hence from the root to leaf nodes, we have a hierarchy of rules involving different numbers of factors. Specifically, since the collection of trees $\{T_1, T_2, \ldots, T_M \mid T_m = (V_m, E_m)\}$ are derived, the corresponding path from the root in each tree gives a rule $R(\cdot)$. Hence, for a given decision tree T_m, the number of generated rules equals $|V_m| - 1$. Such a hierarchy of derived rules naturally incorporates the potential interactions among features. The procedure of decision tree construction also accommodates the data issues when we need to analyze different data types of measurements with missing values.

However, due to many heuristics utilized to ensure the efficiency of random forest, a large number of weakly predictive trees and rules can be generated in this first stage. In order to better select non-redundant and significant rules, we further apply a regularized parametric model to help pruning rules to select a minimum subset of significant rules as the final risk-predictive patterns.

3.3.2 Rule Pruning

For the next rule pruning stage, we apply the overlapped group LASSO regularization to help incorporate the dependency of derived rules when pruning rules to identify significant ones as final risk-predictive patterns. Instead of directly applying the LASSO regularization as in RuleFit, which simply adopts the l_1-norm penalty to produce a sparse coefficient vector with many elements being zero, corresponding to potential weakly predictive or false positive rules, we take an overlapped group LASSO regularization term that allows to enforce considering the dependency relationships among rules when pruning. It is easy to see that when we consider the dependency structure, the overlapped groups can be represented by a hyper-graph with each vertex representing a group of dependent rules and vertices connected by an edge if there are common rules within different groups. In that sense, overlapped group LASSO is a general graph-based regularization method. By incorporating such general dependency structures, we hope to achieve more stable rule discovery by selecting rules with repeatedly selected features. Due to the rule generation procedure, one feature can appear in different rules. Hence, when we incorporate such rule dependency based on common features, the group dependency relationships may lead to overlapped groups. The overlapped group LASSO formulation for rule pruning is therefore critical to consider complex and high-order interactions among original features.

In this rule pruning stage, we first integrate all the rules obtained from the rule generation stage as the predictors transformed from the original features. Given a binary outcome response Y, for example, denoting the corresponding sample coming from healthy or disease subjects, we can formulate a logistic regression problem with the overlapped group LASSO regularization to select significant rules from this large number of random forest rules. Note that for each corresponding rule $R(\mathbf{x})$ as the transformed predictor, it now becomes a binary variable. This enables seamless data integration even when the features takes different types of values. For the pool of all the random forest rules, we can denote them by $R = [R_1, R_2, \ldots, R_N]$, we can now formulate the logistic regression with overlapped group LASSO to fit high-dimensional R against the outcome Y. Similar as the problem (2.13), the formulation for our rule pruning problem is

$$argmin_{\beta,\mu} \|Y - \mu \cdot \mathbf{1} - \mathbf{R}\beta\|_2^2 + \lambda_1 \sum_{j=1}^{J} \sqrt{p_j} \|\beta_j\|_2 + \lambda_2 \|\beta\|_1. \quad (3.1)$$

In the logistic regression problem, we denote the outcome $y_i \in \{-1, 1\}$, and feature vector $\mathbf{x}_i$ for i-th sample, then it becomes

$$argmin_{\beta,\mu} \frac{1}{N} \sum_{i=1}^{N} (\log(1 + \exp(-y_i(\beta^T \mathbf{x}_i + \mu)))) + \lambda_1 \sum_{j=1}^{J} \sqrt{p_j} \|\beta_j\|_2 + \lambda_2 \|\beta\|_1. \quad (3.2)$$

Here $\beta_{\mathbf{j}}$ stands for a subset vector of the elements in β, where $\{\beta_n \mid R_n$ assigned to a given group $j\}$. The logistic loss term, $\frac{1}{N} \sum_{i=1}^{N} (\log(1 + \exp(-y_i(\beta^T \mathbf{x}_i + \mu))))$, is applied to measure prediction accuracy. The l-1 norm penalty $\|\beta\|_1$, is used to minimize the size of selected rule subset. The l-2 norm penalty with respect to each potential group of rules, is used to regularize the diversity within given groups. Here, λ_1 and λ_2 are two parameters to be specified to balance within the three terms, reflecting model accuracy, model complexity, and model internal dependency. With such a formulation, we can implement the recent optimization algorithm [24] to solve (3.1) for rule pruning.

3.3.3 How We Form Overlapped Groups

Deciding the overlapped group structure for rule pruning is critical. Based on the desired rule dependency structures, we propose an overlapped group hyper-graph by grouping rules based on their involvement with both baseline and dynamic features, so that the rules within the same group tend to be selected or pruned together based on their synergistic predictive power on outcome of interest.

First, we consider the rule dependency by assigning rules associated with shared features to the same groups (grouping rules by involved features). Interacting rules (involving two or more features) are assigned to multiple groups and the number of groups will be upbounded by the number of original features. By doing that, in addition to identifying significant rules as risk-predictive patterns, we also can have better interpretation which candidate risk factors may have larger influence on disease progression so that better prognosis and intervention can be designed based on that.

When considering heterogeneous longitudinal data, one of our basic assumptions is to consider the potential temporal dependency from longitudinal measurements. Denote the p-th dynamic feature's measurement collected at the q-th time point by $X_{(p,q)}$. In addition to the rule dependency based on the feature sharing, we also consider the potential interaction of dynamic features at the aligned time points. Such rule dependency may help reveal the interactive effects on disease progression from environmental exposures for example. In this thesis, we only consider grouping rules if they involve dynamic features at the same time point but this can be extended to incorporate potential delays by setting up a time window.

Let's denote a rule that involves both baseline and dynamic features as $R_n \sim (X_{f_1}, \ldots, X_{f_a}, X_{(p,q)_1}, \ldots, X_{(p,q)_b})$. Here, X_{f_a} denotes the a-th baseline feature; and $X_{(p,q)_b}$ denotes the b-th dynamic feature. Considering the previously described rule dependency structures, there are three ways to assign rules into the same group: 1) Given baseline/static feature f fixed, group all the rules $R_n \sim (X_f, \ldots)$ that involve f; 2) Given dynamic feature p, group all the rules $R_n \sim (X_{p,q}, \ldots)$ that involve p; 3) Given time point q for the corresponding longitudinal measurement, group all the rules $R_n \sim (X_{p,q}, \ldots)$ that involve the longitudinal measurements at the same time point q for any dynamic features. We solve (3.2) for rule pruning based on such rule dependency structures modeled in the first term of (3.2).

4 EXPERIMENT RESULTS

In this chapter, we present our exploratory results by applying the structured sparse rule discovery to identify risk-predictive patterns by analyzing a large-scale dataset collected in a prospective T1D study. T1D have been commonly considered as complex disease for subjects with pre-disposed genetic risk. However, the exact trigger(s) and disease progression processes have not been completely understood. In order to identify potential genetic and early-life environmental risk factors to help predict and treat T1D, there have been several large T1D studies to collect diverse types of measurements for candidate risk factors so that systematic analysis of the collected measurements, hoping to discover risk-predictive patterns that may help predict the clinical T1D onset.

4.1 Risk-Predictive Rules for T1D

We focus on a de-identified T1D dataset with heterogeneous measurements, collected for binary, categorical, and numerical valued candidate factors. Some early-life environmental factors, including infectious disease exposures and body-mass index (BMI), have been tracked every 3 months for the first two years for after-born children. In this study, we have 7,512 subjects and we consider whether the subject has been tested positive for persistent auto-antibody tests as the disease outcome of interest. The persistent auto-antibody has been considered as the precursor of the clinical T1D onset. Hence, the early accurate prognosis may help timely intervention to delay and even prevent T1D onset. Among 7.512 subjects in this study, there are 599 subjects with persistent confirmed positive auto-antibodies as the case population and 6,913 negative subjects as the control population. After removing redundant measurements in the original dataset, there are 97 baseline and dynamic candidate factors as features in the analysis. The details of these features are given in a table in Appendix. Among these features, the baseline features are mostly collected from genetic tests, clinic records, daily-life behavior questionnaire, and family information. There are binary features, such as gender, and FDR (first-degree relative) that indicates whether the subject has diabetic first-degree relatives (parents, siblings, etc.); categorical features (race, country, HLA (Human Leukocyte Antigen) genotype category, etc.); and numerical features (mom's age when giving birth, number of days with a specific type of infectious disease, the age of the subject was given certain types of diet, etc.). Dynamic features include BMI and infection history of 4 categories of infectious diseases for the first 2 years of life. For example, the feature ID *inf_epi_group_1_15* denotes how many days the subject has reported (by parents) to be infected with category one infectious diseases on the 15th month.

As described in the previous chapter, we first generate rules based on orginal dataset using random forest package developed and maintained by A. Liaw, *et al.*[25], which naturally incorporates potential interactions among different features. From the adopted random forest procedure, we have generated 952 candidate rules by setting the number of trees to 250, and maximum node number for each tree as 4. The random forest training takes 9.435s. Besides, by having the assumption that there are a few features ($HLA\_Category$, Sex, and FDR) being conjectured to be risk-predictive and have significant marginal effects, we want to validate by creating additional rules just based on taking every possible value of each of those features. For example, the additional rules for FDR are $FDR \in \{'1'\}$ and $FDR \in \{'0'\}$, for feature Sex are $Sex \in \{'Male'\}$ and $Sex \in \{'Female'\}$, and for $HLA\_Category$ we have 12 more rules indicating whether subject belongs to particular HLA category. Thus, we have 16 more rules, finally 968 rules in total. Then, we select rules with the overlapped group LASSO formulation by pruning away weakly predictive or redundant rules. Again, the rule dependency structures are considered by grouping rules in three ways: 1) if the rules involve the same baseline feature; 2) if the rules involve different dynamic features at same time point($e.g$ bmi_03 and inf_epi_group_2_03), or 3) if the rules involve the same dynamic features, even at different time points ($e.g$ bmi_03 and bmi_09).

When solving the overlapped group LASSO logistic regression, we make use of the SLEP package developed by J. Liu, *et al.*[26], which solves overlapped group LASSO efficiently. And grid search has been implemented to search for the optimal penalty coefficients for for the model complexity penalty l-1 norm, and for the overlapped group penalty l-2 norm. We search for the appropriate λ_1 value between $[0.001, 0.05]$ with the step size 0.001, jointly with λ_2 between $[0.0001, 0.005]$ with the step size 0.0001, totally $50 \times 50 = 2500$ setups, repeated 10 times each. The average running time of every 10 repeated experiments is 1.073s. Based on the average prediction accuracy, AUC (Area Under Curve) of Receiver Operating Characteristic (ROC), as well as the number of non-zero coefficients obtained from the exhaustive search from neighboring grids in the parameter space, we set $\lambda_1 = 0.012$ and $\lambda_2 = 0.0023$ in (3.2).

Ranked by the absolute value of the derived model coefficients in overlapped group LASSO logistic regression, the top 15 rules are provided in Table 4.1.

In this table, positive coefficient values reveal that the corresponding rules characterize the subpopulation that may have increased risk of developing T1D while negative coefficients characterize the protective patterns. The larger the magnitudes or absolute values of the corresponding rules are; the greater the corresponding features have impact on T1D development. We also report both the Support and Important indices for the identified rules. The Support index indicates how many percentages of the population are the samples that satisfy the corresponding rules. For example, if Support equals to 0.5, then the corresponding rule divides the set into halves: 50% satisfying the rule and the other 50% violating the rule. The Importance index is derived based on both the corresponding model coefficient and Support, see (2.7).

Table 4.1: The Top 15 Identified Rules from OGL

Rank	Description	Coefficient	Support	Importance
1	$HLA\_Category \in \{'9'\}$	-0.004947	0.2031	0.0020
2	$HLA\_Category \in \{'1'\}$	0.004713	0.3857	0.0023
3	$HLA\_Category \in \{'1','2','6','8'\}$ & $bmi\_00 > 12.715$	0.0046	0.4270	0.0023
4	$HLA\_Category \in \{'1','2','4','6','7','8'\}$ & $potatoes > 23$	0.0046	0.7553	0.0020
5	$FDR \in \{'1'\}$	0.0044	0.1138	0.0014
6	$HLA\_Category \in \{'1','2','4','5','6','7'\}$ & $rice\_milk > 442$	0.0043	0.6850	0.0020
7	$HLA\_Category \in \{'1','5','6'\}$ & $barley \leq 760.5$	0.0043	0.3946	0.0021
8	$HLA\_Category \in \{'3','4','7','9','10'\}$ & $cereals > 10.5$	-0.0043	0.3859	0.0021
9	$FDR \in \{'0'\}$ & $common\_coldTotal \leq 27.5$	-0.0039	0.7941	0.0016
10	$HLA\_Category \in \{'1','5','6','8'\}$ & $cereals \leq 323.5$	0.0038	0.4153	0.0019
11	$subj\_curr\_age > 6.5$ & $babysweightgrams > 3307.5$	0.00375	0.3926	0.0018
12	$FDR \in \{'0'\}$ & $bmi\_12 \leq 18.86$	-0.0035	0.7718	0.0015
13	$HLA\_Category \in \{'3','4','7','9','10'\}$ & $common\_coldTotal \leq 30.50$	-0.0033	0.3606	0.0016
14	$FDR \in \{'0'\}$ & $bmi\_06 \leq 20.10$	-0.0032	0.8381	0.0012
15	$common\_coldTotal \leq 12.96$	-0.0029	0.5451	0.0014

4.2 Rule Validation via Survival Analysis

To further validate the identified rules, we carry out survival analysis on these derived rules to estimate the separability of two subpopulations distinguished by the corresponding rule. Since we have the clinical T1D diagnosed date for each subject (*t1d_diag_date*) for who eventually gets diagnosed with T1D clinical onset, we use these records to plot Kaplan-Meier (KM) curves for every two subpopulations partitioned by each top rule. The KM plots for the top 15 rules are given in Figures 4.1, 4.2, 4.3, 4.4, 4.5, 4.6, 4.7, 4.8, 4.9,

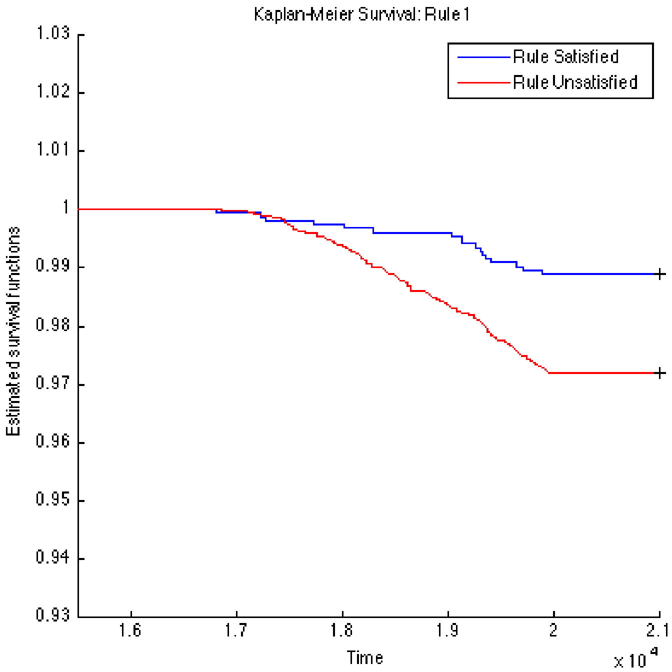


Figure 4.1: K-M: Rule01 OGL

4.10, 4.11, 4.12, 4.13, 4.14, and 4.15. The horizontal axis in the KM plot represents the aligned time since the birth of subjects, and the vertical axis represents the survival rate (the percentage of those have not been diagnosed with T1D) for a given subpopulation. The stared curves indicate the rule-satisfied subpopulations, and the cycled curves indicate the rule-violated ones.

It is clear that all these 15 top rules are risk-predictive, especially Rules 5, 9, 12, and 14. When applying log-rank tests to compare the survival distributions of subpopulations under individual rules, all the identified rules are significant with p-values < 0.05 as illustrated in Table 4.2. When we check Rules 5, 9, 12, and 14, there are both risk-increasing and risk-protective patterns with their log-rank test p-values $< 10^{-9}$. This group of rules involve mostly the risk factors FDR, infection history, BMI and their interactions. These indeed can be found reported evidences in the T1D literature.

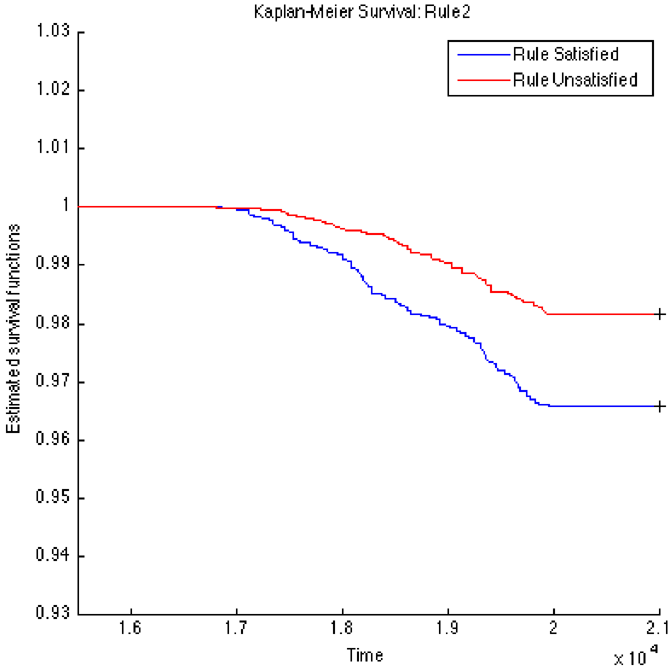


Figure 4.2: K-M: Rule02 OGL

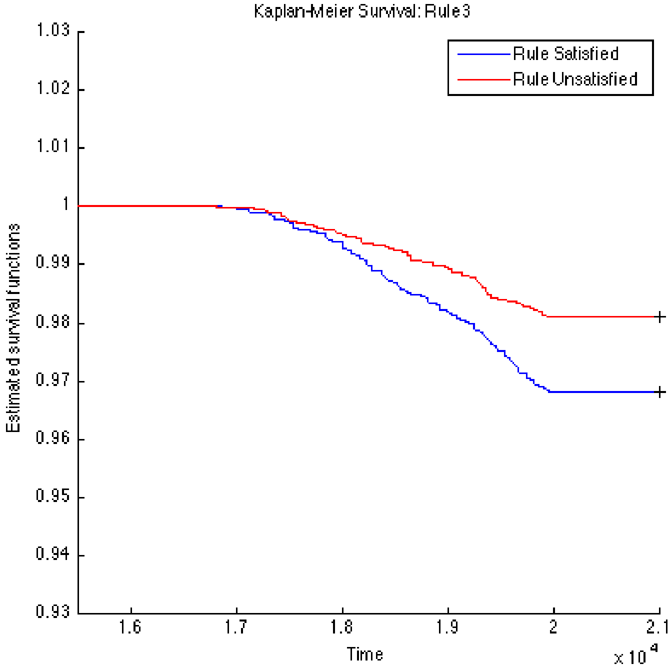


Figure 4.3: K-M: Rule03 OGL

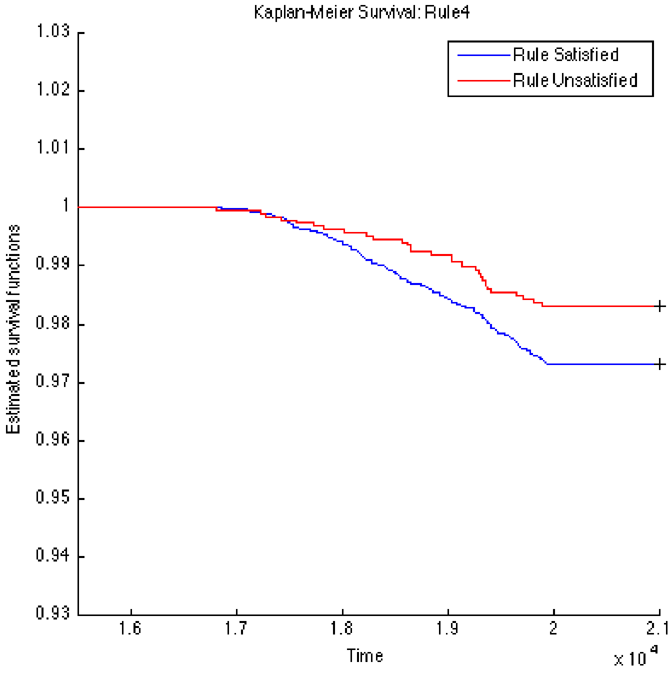


Figure 4.4: K-M: Rule04 OGL

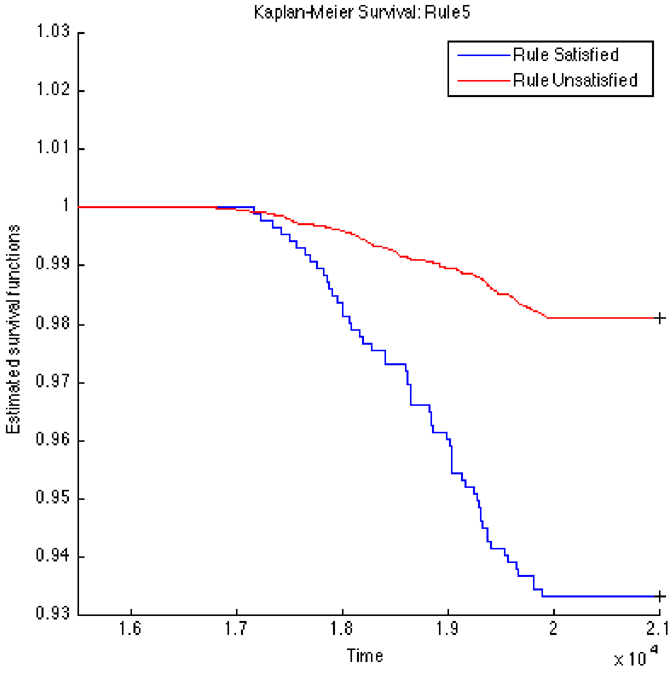


Figure 4.5: K-M: Rule05 OGL

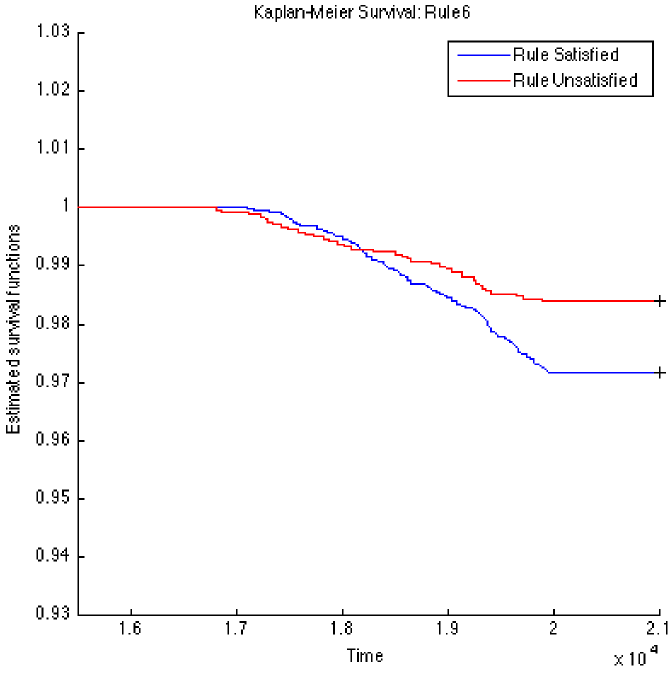


Figure 4.6: K-M: Rule06 OGL

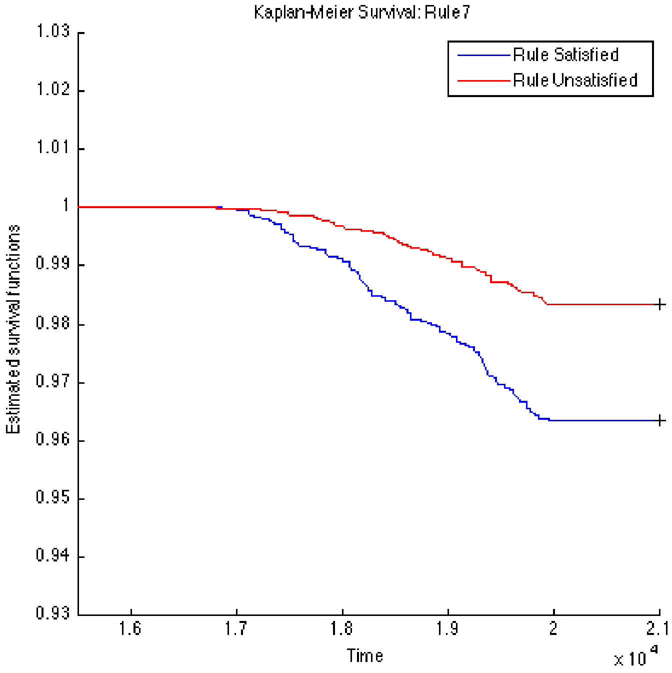


Figure 4.7: K-M: Rule07 OGL

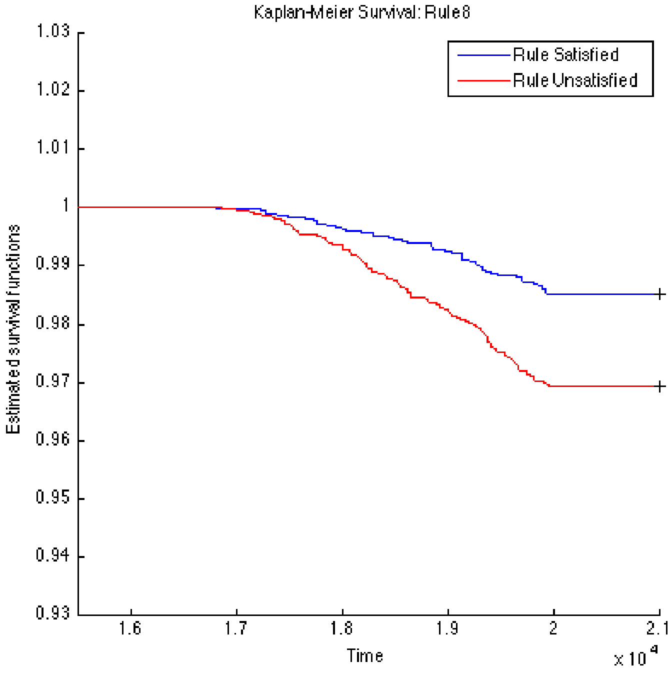


Figure 4.8: K-M: Rule08 OGL

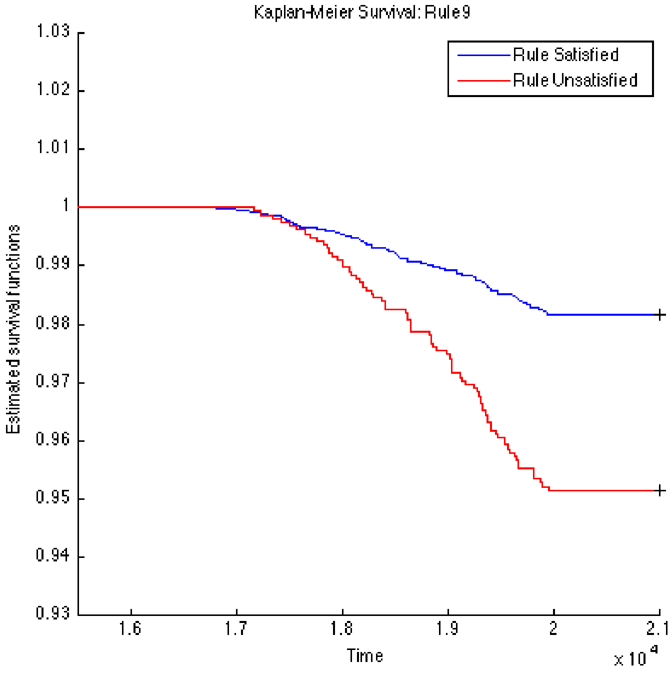


Figure 4.9: K-M: Rule09 OGL

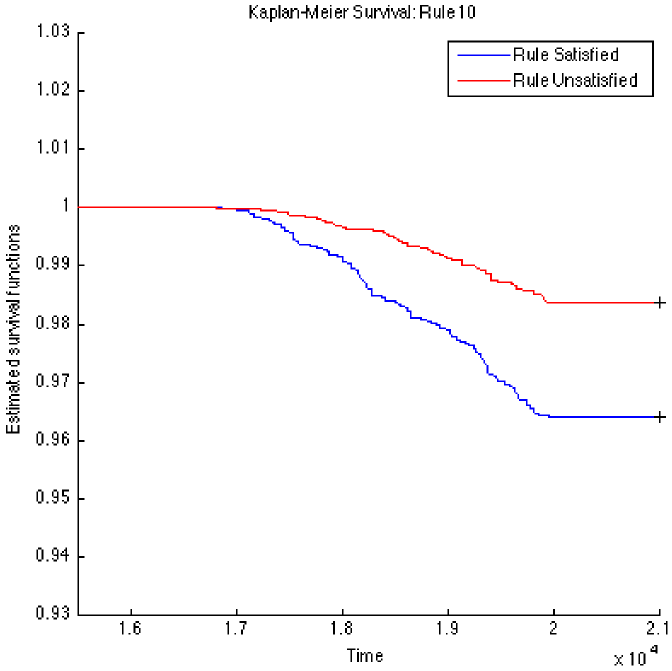


Figure 4.10: K-M: Rule10 OGL

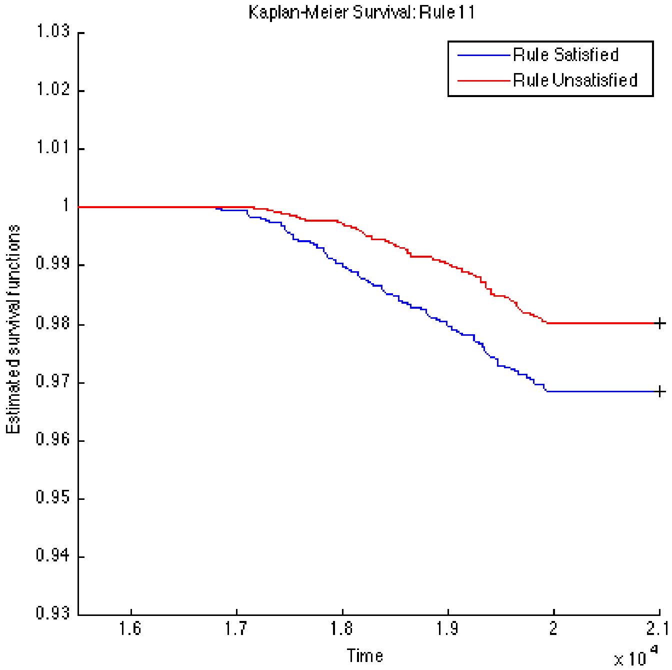


Figure 4.11: K-M: Rule11 OGL

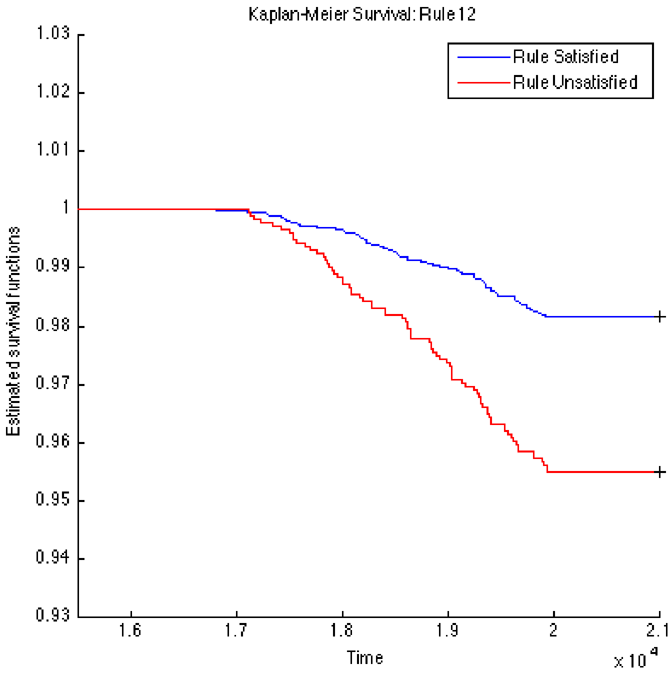


Figure 4.12: K-M: Rule12 OGL

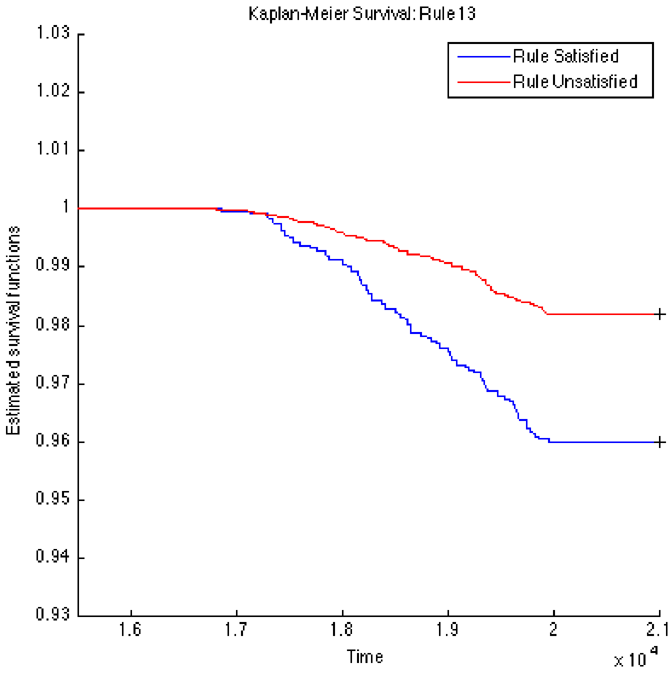


Figure 4.13: K-M: Rule13 OGL

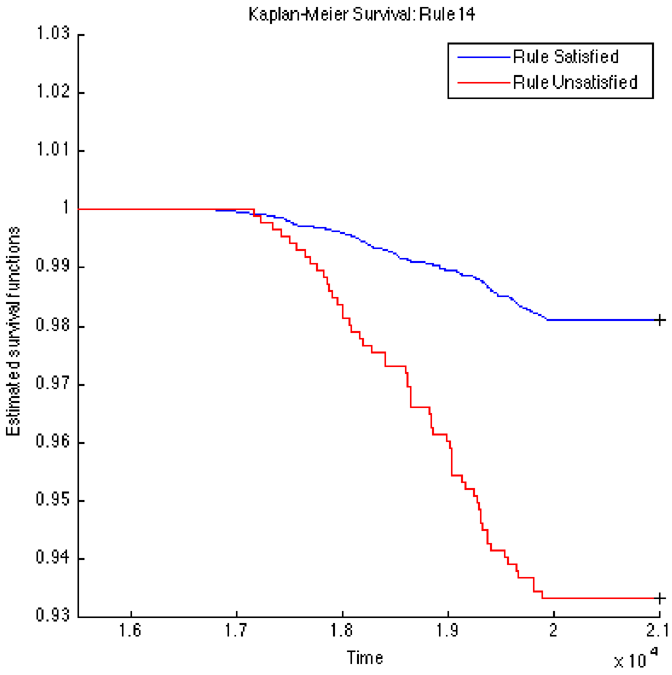


Figure 4.14: K-M: Rule14 OGL

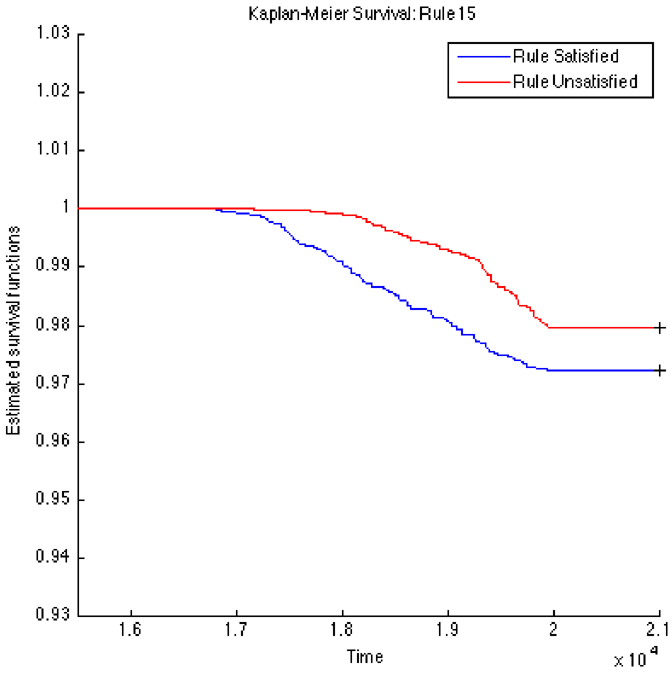


Figure 4.15: K-M: Rule15 OGL

Table 4.2: Log-Rank Test p-values of Top 15 Rules from OGL

Rules	P-value of log-rank test	Rules	P-value of log-rank test
Rule 1	$2.3260e-4$	Rule 9	$1.2095e-11$
Rule 2	$2.2292e-5$	Rule 10	$9.7659e-8$
Rule 3	$5.3404e-4$	Rule 11	$1.8051e-3$
Rule 4	$1.8996e-2$	Rule 12	$6.8077e-10$
Rule 5	$< 1e-16$	Rule 13	$5.9548e-5$
Rule 6	$1.8573e-3$	Rule 14	$2.9132e-13$
Rule 7	$8.8449e-8$	Rule 15	$4.3533e-2$
Rule 8	$2.4628e-5$		

4.3 Comparison to RuleFit with LASSO Rule Pruning

To further evaluate the rules selected from overlapped group LASSO and compare them with the results from RuleFit, we have identified top 15 rules via LASSO (without considering the overlapped group penalty term $\lambda_2 = 0$) as in RuleFit. The corresponding penalty parameter λ_1 for the model complexity penalty in (3.2) in this experiment is set to 0.03, which is also obtained from exhaustive grid search within an interval $[0.0001, 0.005]$. The search step size was set to 0.0001. We have run evaluations 10 times for each setup, to compare the average obtained prediction accuracy, non-zero coefficients and AUC values. The average running time of every 10 repeated experiments is 0.792s, also using SLEP package[26]. With $\lambda_1 = 0.03$, the logistic regression model with the same set of random forest rules yields a decent prediction accuracy and selects a reasonable number of non-zero model coefficients. Here we list the top 15 rules from LASSO rule pruning in Table 4.3, as a comparison to the rules selected in our overlapped group LASSO framework.

Table 4.3: The Top 15 Identified Rules from LASSO

Rank	Description	Coefficient	Support	Importance
1	$common\_coldTotal \leq 27.5$ & $FDR \in \{'0'\}$	-0.0161	0.7940	0.0065
2	$HLA\_Category \in \{'3','4','7','9','10'\}$ $common\_coldTotal \leq 30.50$	-0.0095	0.3606	0.0045
3	$HLA\_Category \in \{'1','2','4','5','6','7'\}$ & $rice\_milk > 442$	0.0086	0.6849	0.0040
4	$HLA\_Category \in \{'2','3','4','7','8','9','10'\}$ & $bmi\_00 \leq 19.52$	-0.0072	0.5836	0.0036
5	$HLA\_Category \in \{'1','5','6','8'\}$ & $cereals \leq 323.5$	0.0073	0.4153	0.0036
6	$HLA\_Category \in \{'1','2','6','8'\}$ & $bmi\_00 > 12.715$	0.0071	0.4269	0.0035
7	$HLA\_Category \in \{'1','6','8'\}$ & $milk\_product \leq 525$	0.0068	0.3908	0.0033
8	$HLA\_Category \in \{'1','5','6'\}$ & $barley \leq 525$	0.0067	0.3946	0.0033
9	$HLA\_Category \in \{'2','3','4','5','7','9','10'\}$ & $bmi\_12 \leq 18.864$	-0.0067	0.5213	0.0033
10	$HLA\_Category \in \{'2','3','4','7','9','10'\}$ & $inf\_epi\_2\_21 \leq 323.5$	-0.0066	0.5821	0.0033
11	$HLA\_Category \in \{'2','3','4','5','7','8','9','10'\}$ & $rice > 10.5$	-0.0054	0.6021	0.0027
12	$FDR \in \{'0'\}$ & $inf\_epi\_3\_18 \leq 0.5$	-0.0052	0.8272	0.0020
13	$HLA\_Category \in \{'1'\}$	0.0048	0.3857	0.0023
14	$HLA\_Category \in \{'1','5','6'\}$ & $bmi\_18 > 13.41$	0.0073	0.4153	0.0036
15	$HLA\_Category \in \{'1','6'\}$ & $bmi\_21 > 12.98$	0.0042	0.3850	0.0020

As we can see from Tables 4.1 and 4.3, most selected rules by overlapped group LASSO remain consistent with those from LASSO by RuleFit. Two collections of selected rules typically involve similar features and cutoffs ($e.g.$ $HLA\_Category \in \{'1','5','6'\}$ indeed is known as the genotypes with higher risk of developing T1D). We also have observed early-life environmental factors including BMI and infection history also have interactive effects on T1D disease development.

We also have performed survival analysis with KM plots as well as log-rank tests on

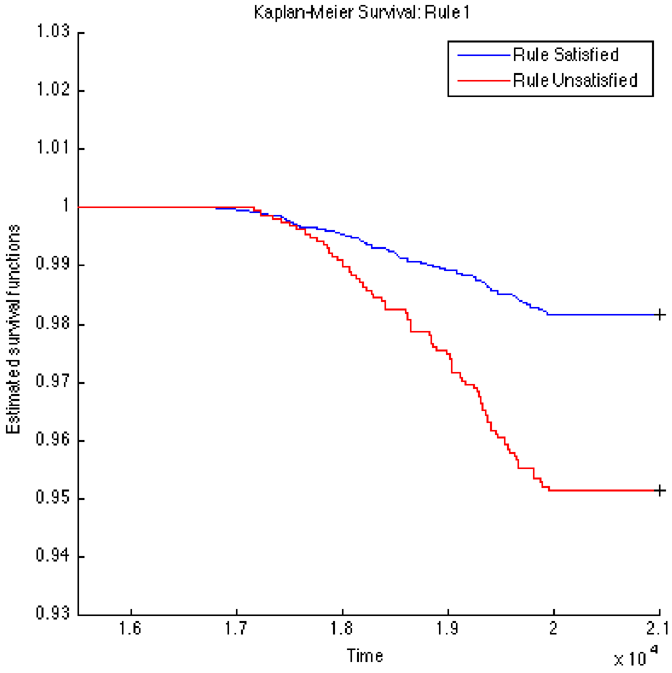


Figure 4.16: K-M: Rule01 LASSO

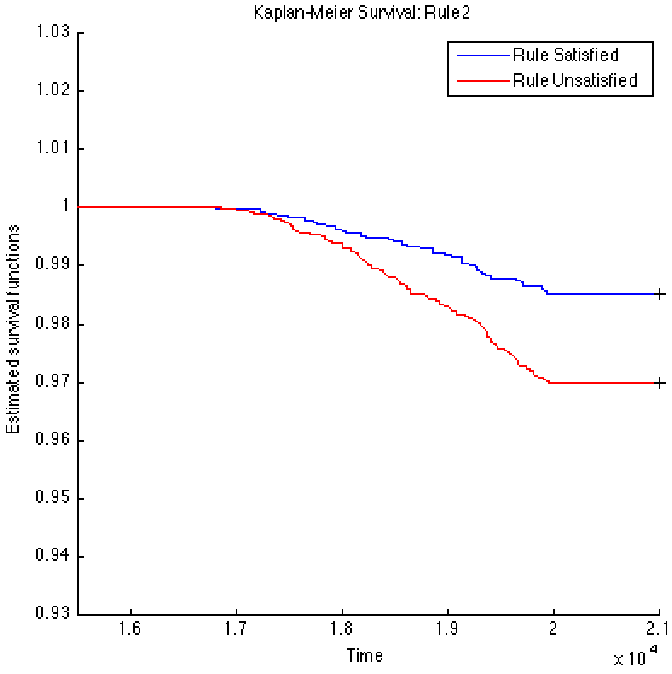


Figure 4.17: K-M: Rule02 LASSO

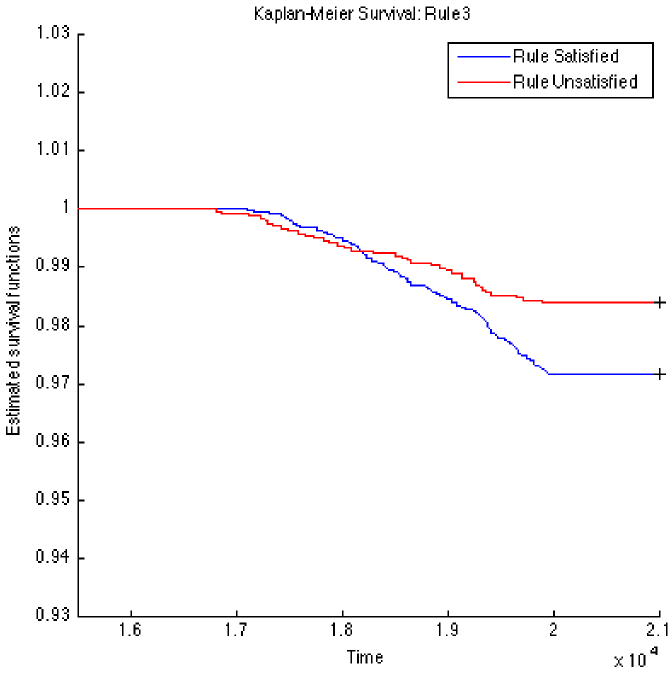


Figure 4.18: K-M: Rule03 LASSO

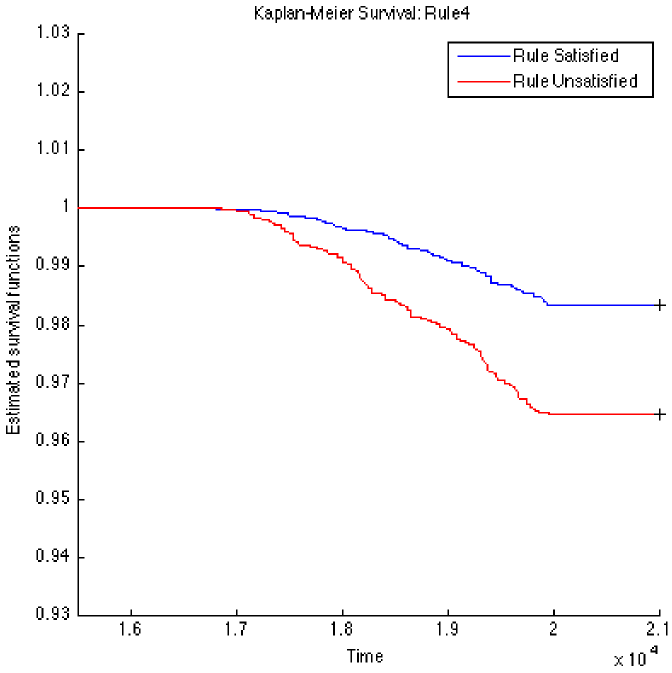


Figure 4.19: K-M: Rule04 LASSO

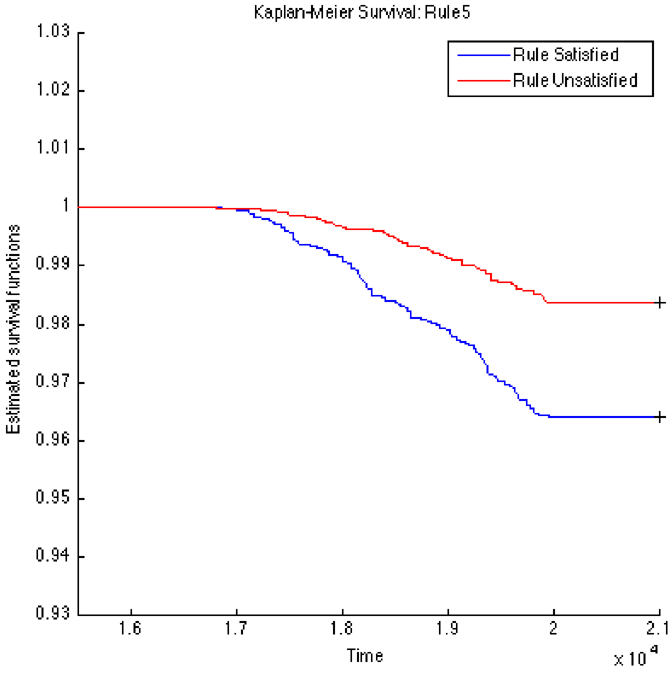


Figure 4.20: K-M: Rule05 LASSO

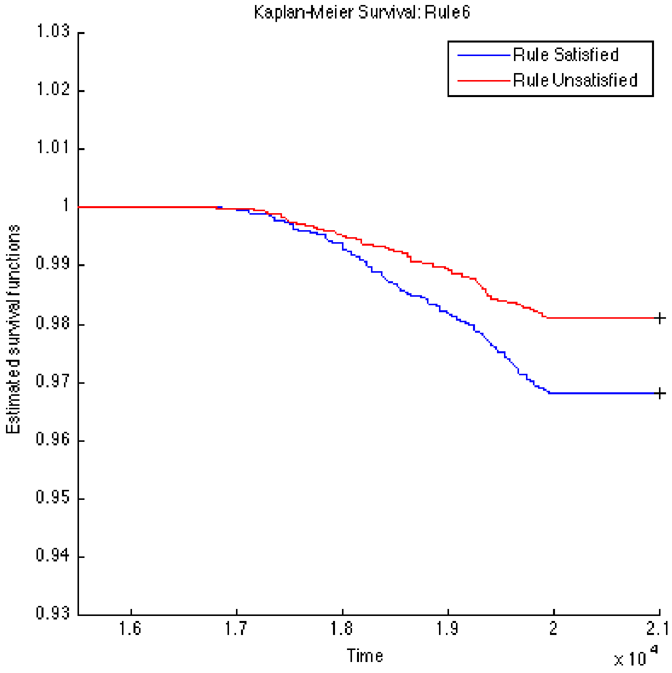


Figure 4.21: K-M: Rule06 LASSO

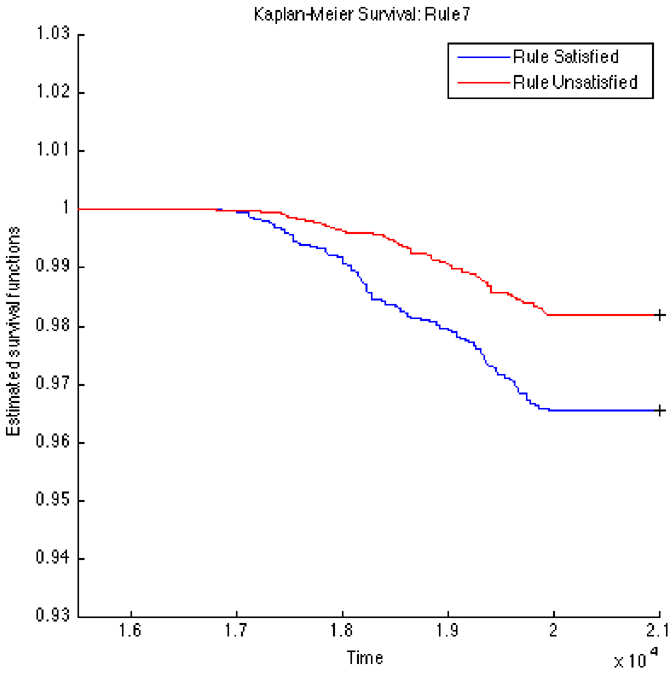


Figure 4.22: K-M: Rule07 LASSO

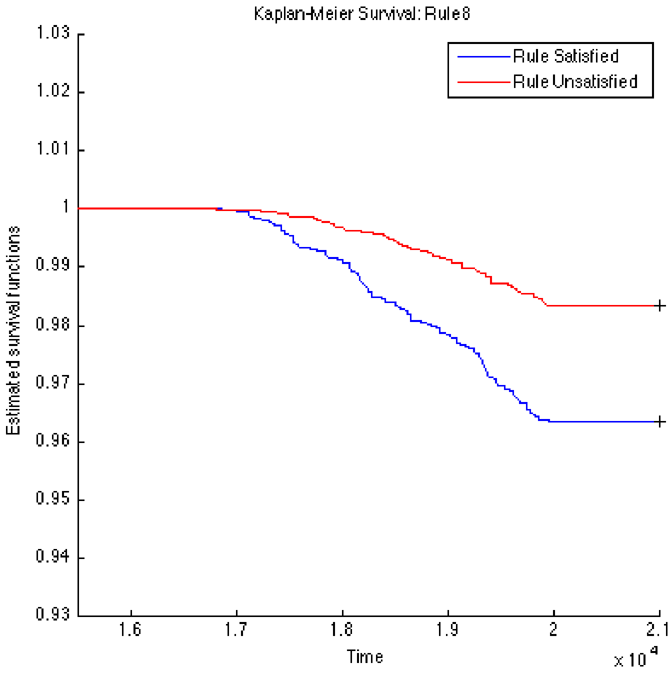


Figure 4.23: K-M: Rule08 LASSO

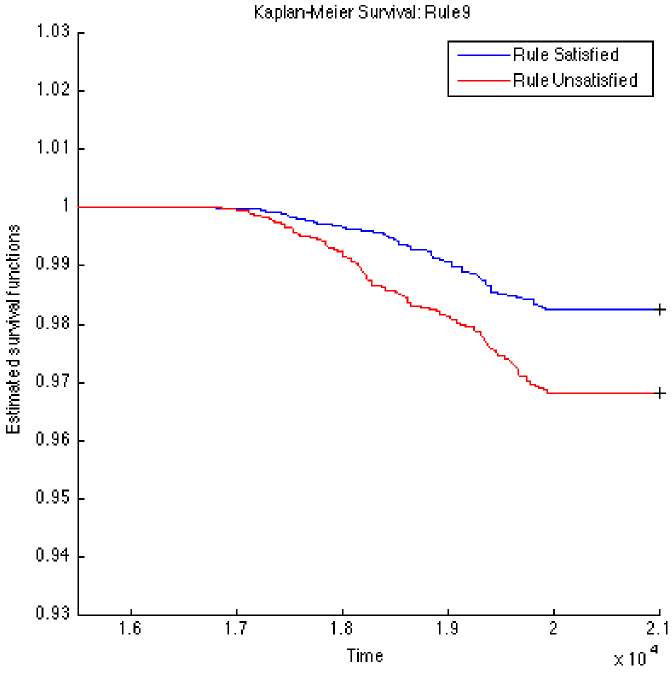


Figure 4.24: K-M: Rule09 LASSO

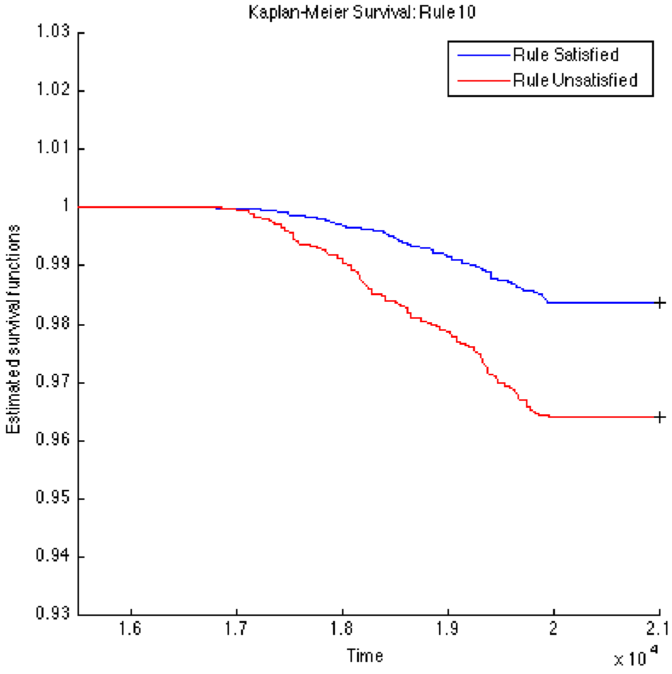


Figure 4.25: K-M: Rule10 LASSO

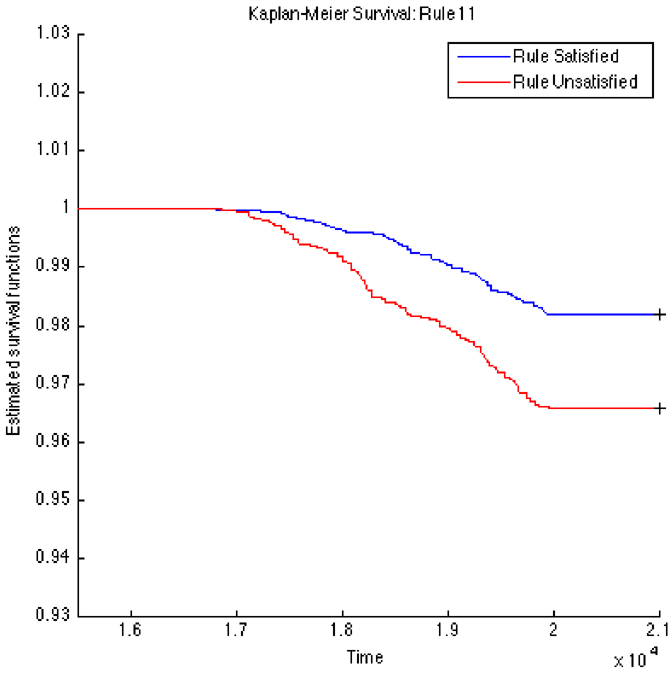


Figure 4.26: K-M: Rule11 LASSO

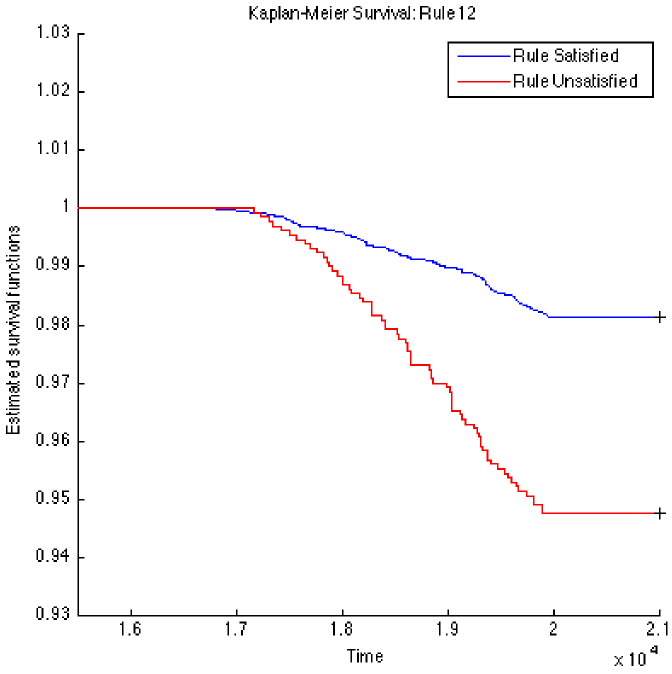


Figure 4.27: K-M: Rule12 LASSO

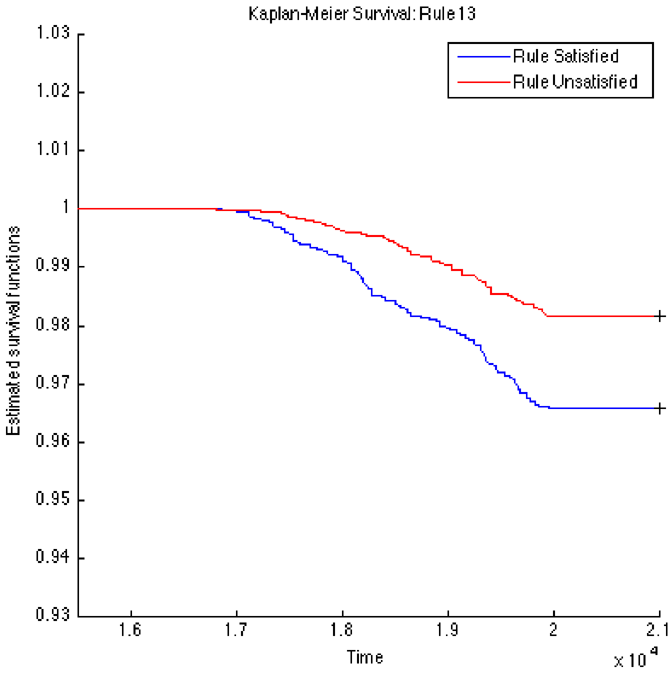


Figure 4.28: K-M: Rule13 LASSO

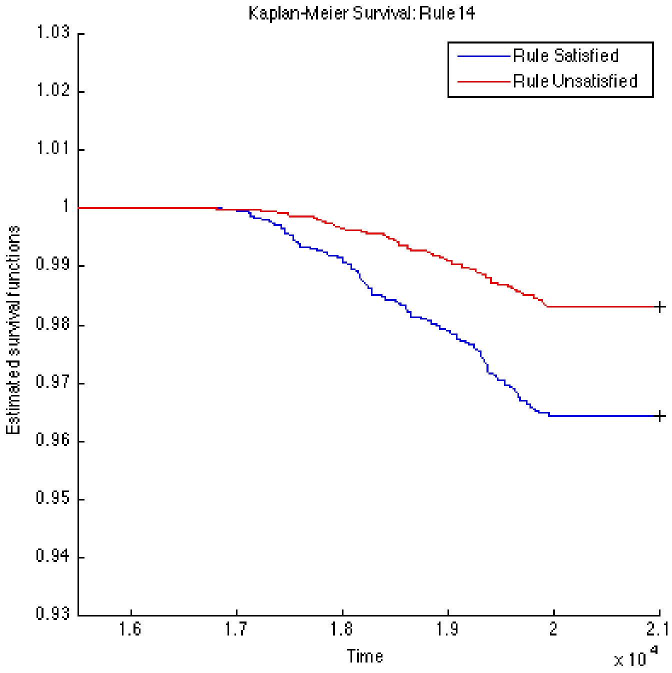


Figure 4.29: K-M: Rule14 LASSO

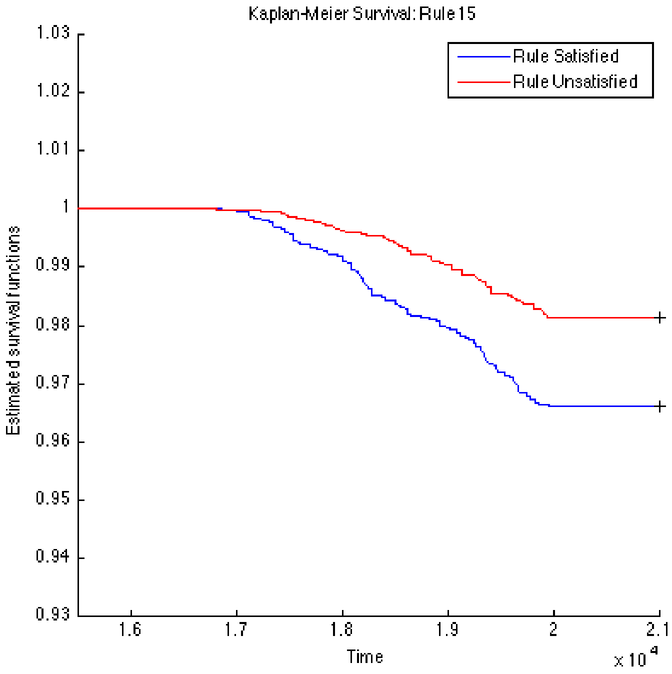


Figure 4.30: K-M: Rule15 LASSO

Table 4.4: Log-Rank Test *p*-values of Top 15 Rules from LASSO

Rules	P-value of log-rank test	Rules	P-value of log-rank test
Rule 1	$1.2095e-11$	Rule 9	$7.6264e-5$
Rule 2	$5.9548e-5$	Rule 10	$6.3691e-8$
Rule 3	$1.8573e-4$	Rule 11	$1.4774e-5$
Rule 4	$2.6483e-7$	Rule 12	$1.2430e-12$
Rule 5	$9.7659e-8$	Rule 13	$2.2292e-5$
Rule 6	$5.3404e-4$	Rule 14	$4.7007e-7$
Rule 7	$1.1237e-5$	Rule 15	$3.9779e-5$
Rule 8	$8.8450e-8$		

the rules identified with LASSO, as illustrated in Figures 4.16, 4.17, 4.18, 4.19, 4.20, 4.21, 4.22, 4.23, 4.24, 4.25, 4.26, 4.27, 4.28, 4.29, 4.30, and Table 4.4. Clearly selected rules can serve as risk-predictive patterns for T1D development.

The rules highly ranked in LASSO approach and overlapped group LASSO approach show strong consistency, as similar risk-predictive features are involved, such as $HLA\_Category$, $common\_coldTotal$, FDR, and BMI. Involved features also exhibit similar predictive patterns, *e.g.*, a larger BMI, $HLA\_Category \in \{'1','5','6'\}$ generally indicates higher risk in our model. These discovered pattern turns out to be valid as also reported in the diabetes literature [27, 28].

Comparing to RuleFit with LASSO rule pruning, our structured sparse rule discovery with overlapped group LASSO indeed have the trend to select related factors into the top rules as we noted earlier with Rules 5, 9, 10, and 12 from overlapped group LASSO. These corresponding rules show strong statistical significance in survival analysis; and they are risk-predictive, and worthy to be selected and examined further.

In addition, with overlapped group LASSO, the rule pruning procedure may help reduce false discovery of interacting rules. Comparing the top rules from overlapped group LASSO and LASSO, more rules involving only one factor have been ranked high in overlapped group LASSO due to imposing the rule dependency. It naturally adjusts the contributions of complex interacting rules to reduce the false discovery.

Another trend can also be observed when checking the Support indices in Table 4.3. RuleFit with LASSO pruning tend to given the higher rank to the rules with Support close to 0.5 (which means preferred rules divide the whole set into halves), than those from overlapped group LASSO shown in Table 4.1. In fact, among the top 15 selected rules, the average Support for overlapped group LASSO top rules is 0.4979, with the standard deviation 0.2226, and that for RuleFit rules is 0.5180, with the standard deviation 0.1550. Again, due to the rule dependency enforced by overlapped group LASSO, our structured sparse rule discovery does not restrict to achieving good global splitting only, but also selects the rules that may help distinguish interesting subpopulations.

Since we have observed that some important rules denoting potentially interesting subpopulations are only selected by overlapped group LASSO, we may conclude that, our overlapped group LASSO approach may help find more robust and refined subpopulations for better understanding and prediction of T1D development, compared to the original RuleFit, due to the consideration of the dependency relationships among different features and temporal measurements.

4.4 Prediction Accuracy Evaluation

Our method is not only effective in identifying top rules as disease risk-predictive patterns, but also capable of performing prediction for disease diagnosis. As a typical rule-based ensemble model, in the form of overlapped group LASSO logistic regression, the trained classifier can be applied to derive the overall risk score of a given subject.

We evaluate the prediction accuracy by AUC values. In this evaluation, in order to avoid reporting bias, we perform random sampling to obtain training and testing samples, and repeated training the model for 200 times. Each time training dataset and testing dataset are separately random sampled from original set, both with balanced outcome labels. And we want to keep the training and testing set both balanced, by which the training set is produced with balanced case and control samples, especially because the case samples have a much smaller portion (less than 10%) in the original dataset. For each evaluation experiment, we randomly sample the 63% of all the case samples (since case samples take a much smaller portion) and randomly sample the same number of control samples. Then we use the model trained with the training set to make prediction on the testing set, thus obtained AUC under the current evaluation. The average AUC is obtained by repeating this process, which turns out to be 0.6545.

For comparison, we also have evaluated the LASSO regularized logistic regression [5] by RuleFit, which yields an average AUC value at 0.6570. In addition, we plot

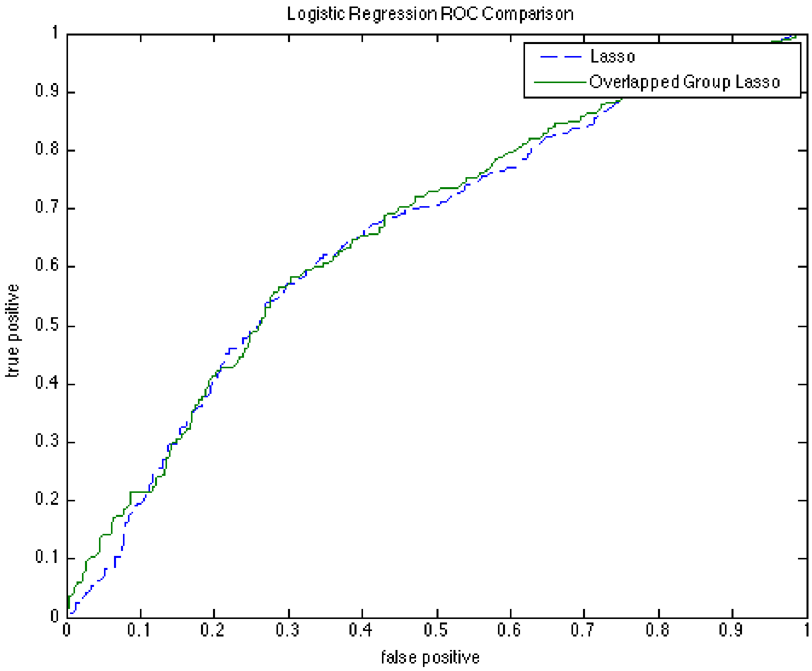


Figure 4.31: ROC Comparison

the ROC curves by our overlapped group LASSO model and RuleFit from one of 200 evaluation experiments, which are very close to each other. It is interesting to see that our proposed structured sparse rule discovery with overlapped group LASSO does not sacrifice too much on prediction accuracy but have the potential to incorporate more complicated dependency structures when selecting top rules as risk-predictive patterns.

We note that our structured rule discovery with overlapped group LASSO achieves slighly worse AUC compared to RuleFit-based analysis [5]. This is reasonable as we impose additional structure dependency regularizations. More critically, our method is designed to discover risk-predictive patterns. We have observed a number of rules can significantly distinguish subgroups of subjects with different trends in survival, and it is consistent with the existing findings in relevant studies. Hence, we believe our proposed structured rule-based method has the potential in helping understand disease progression. In order to comprehensively evaluate the proposed method, we will need to conduct more experiments with well-established benchmark datasets and compare with other existing methods as similarly done in [5].

5 CONCLUSIONS

In this thesis, we have designed a structured rule-based disease risk-predictive pattern discovery framework by extending RuleFit to explicitly model potential rule dependency so that heterogeneous longitudinal data can be appropriately analyzed. Our structured spare rule discovery procedure has a few nice properties comparing to the original RuleFit. By applying overlapped group LASSO, we have the flexibility to model the dependency structures among rules and features. This property is especially useful in a situation where we have some prior knowledge or assumptions about candidate factors.

We have obtained exploratory results by applying both RuleFit and our structured sparse rule discovery with overlapped group LASSO regularization to analyze a heterogeneous dataset studying T1D disease progression. Our experiments have shown that its capability of identifying risk-predictive patterns for T1D development. The identified rules are consistent with the reported T1D research results. For instance, HLA genotypes have been reported as an indicator of susceptibility to T1D [29]. BMI has also been conjectured as a significant indicator of diabetes risk, especially for people at younger ages [28]. FDR is observed in our experiments as one of the dominant risk factors, which is also reported in the literature [27] as an important risk factor of T1D onset.

5.1 Future Directions

For analyzing complex disease longitudinal data, careful modeling temporal dependency can help reveal disease progression mechanisms and timely intervene to help delay and prevent disease onset. It is definitely an important direction for future reseach.

There can be several possible improvements for longitudinal data analysis. From our experimental results, we have not been able to figure out significant time-associated patterns. One possible explanation may be that the temporal dependency with respect to the disease development may not be as strong as the other interactive effects identified in this specific dataset. On the other hand, only the rules with dynamic features having observations at the same time points are grouped but more general temporal dependency and continuity at neighboring time points may need to be considered. For example, the feature *bmi*_03 should be more correlated to *bmi*_06 rather than to *bmi*_24. Such dependency relationships may need to be integrated into the future rule-based discovery models. With the flexibility of overlapped group LASSO, we can establish a complex interaction network by adding more general dependency relationships, which may lead to more interesting patterns and association findings related to disease progression. We will study these potential models in our future research.

REFERENCES

[1] M. A. Brown, L. G. Kennedy, A. J. Macgregor, C. Darke, E. Duncan, J. L. Shatford, A. Taylor, A. Calin, and P. Wordsworth, “Susceptibility to ankylosing spondylitis in twins the role of genes, hla, and the environment,” *Arthritis & Rheumatism*, vol. 40, no. 10, pp. 1823–1828, 1997.

[2] D. A. Di Monte, “The environment and parkinson’s disease: is the nigrostriatal system preferentially targeted by neurotoxins?,” *The Lancet Neurology*, vol. 2, no. 9, pp. 531–538, 2003.

[3] R. Brookmeyer, S. Gray, and C. Kawas, “Projections of alzheimer’s disease in the united states and the public health impact of delaying disease onset.,” *American Journal of Public Health*, vol. 88, no. 9, pp. 1337–1342, 1998.

[4] T. Welborn, M. Knuiman, V. McCann, K. Stanton, and I. Constable, “Clinical macrovascular disease in caucasoid diabetic subjects: logistic regression analysis of risk variables,” *Diabetologia*, vol. 27, no. 6, pp. 568–573, 1984.

[5] Y. Lin, X. Qian, J. Krischer, K. Vehik, H.-S. Lee, and S. Huang, “A rule-based prognostic model for type 1 diabetes by identifying and synthesizing baseline profile patterns,” *PloS one*, vol. 9, no. 6, p. e91095, 2014.

[6] J. H. Friedman and B. E. Popescu, “Predictive learning via rule ensembles,” *The Annals of Applied Statistics*, pp. 916–954, 2008.

[7] J. Friedman, T. Hastie, and R. Tibshirani, “A note on the group lasso and a sparse group lasso,” *ArXiv Preprint ArXiv:1001.0736*, 2010.

[8] B. P. Tabaei and W. H. Herman, “A multivariate logistic regression equation to screen for diabetes development and validation,” *Diabetes Care*, vol. 25, no. 11, pp. 1999–2003, 2002.

[9] J. Friedman, “Rulefit with r,” 2005.

[10] S. R. Safavian and D. Landgrebe, “A survey of decision tree classifier methodology,” 1990.

[11] L. Hyafil and R. L. Rivest, “Constructing optimal binary decision trees is np-complete,” *Information Processing Letters*, vol. 5, no. 1, pp. 15–17, 1976.

[12] L. Breiman, “Technical note: Some properties of splitting criteria,” *Machine Learning*, vol. 24, no. 1, pp. 41–47, 1996.

[13] A. A. Al Jarullah, “Decision tree discovery for the diagnosis of type ii diabetes,” in *Innovations in Information Technology (IIT), 2011 International Conference on*, pp. 303–307, IEEE, 2011.

[14] C. W. Olanow, R. L. Watts, and W. C. Koller, “An algorithm (decision tree) for the management of parkinson’s disease (2001): treatment guidelines,” *Neurology*, vol. 56, no. suppl 5, pp. S1–S88, 2001.

[15] Y. Freund, R. E. Schapire, *et al.*, “Experiments with a new boosting algorithm,” in *Icml*, vol. 96, pp. 148–156, 1996.

[16] L. Breiman, “Bagging predictors,” *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[17] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[18] M. Khalilia, S. Chakraborty, and M. Popescu, “Predicting disease risks from highly imbalanced data using random forest,” *BMC Medical Informatics and Decision Making*, vol. 11, no. 1, p. 1, 2011.

[19] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[20] M. Haghighi, S. B. Johnson, X. Qian, K. F. Lynch, K. Vehik, S. Huang, T. S. Group, *et al.*, “A comparison of rule-based analysis with regression methods in understanding the risk factors for study withdrawal in a pediatric study,” *Scientific Reports*, vol. 6, 2016.

[21] H. Zou and T. Hastie, “Regularization and variable selection via the elastic net,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[22] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, “Sparsity and smoothness via the fused lasso,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 1, pp. 91–108, 2005.

[23] M. Yuan and Y. Lin, “Model selection and estimation in regression with grouped variables,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.

[24] L. Yuan, J. Liu, and J. Ye, “Efficient methods for overlapping group lasso,” in *Advances in Neural Information Processing Systems*, pp. 352–360, 2011.

[25] A. Liaw and M. Wiener, “Classification and regression by randomforest,” *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[26] J. Liu, S. Ji, and J. Ye, *SLEP: Sparse learning with efficient projections*. Arizona State University, 2009.

[27] B. Muktabhant, P. Sanchaisuriya, M. Trakulwong, R. Mingchai, and F. P. Schelp, “A first-degree relative with diabetes mellitus is an important risk factor for rural thai villagers to develop type 2 diabetes mellitus,” *Asia-Pacific Journal of Public Health*, p. 1010539514555861, 2014.

[28] K. V. Narayan, J. P. Boyle, T. J. Thompson, E. W. Gregg, and D. F. Williamson, “Effect of bmi on lifetime risk for diabetes in the us,” *Diabetes Care*, vol. 30, no. 6, pp. 1562–1566, 2007.

[29] Y. Park, C. Wang, K. Ko, S. Yang, M. Park, M. Yang, and J.-X. She, “Combinations of hla dr and dq molecules determine the susceptibility to insulin-dependent diabetes mellitus in koreans,” *Human Immunology*, vol. 59, no. 12, pp. 794–801, 1998.

APPENDIX A
MISCELLANEOUS

A.1 Figures/Tables in Appendix

A.1.1 Feature Dictionary

Table A.1: Feature Description

Feature Name	**Description**
BabyBirthType	Baby born type
total_roots	Age at first introduction, in days, of any roots
child_first_start	age in weeks supplement started
pork_beef	Age at first introduction, in days, of any pork or beef
subj_curr_age	Subject current age in years
suppl_flag	If subject has taken supplements
noncorn_cereals	Age at first intro to any cereal excluding corn
rye	Age at first introduction, in days, of any rye
regular_cow_s_milk_or_ice_cream	Age at first intro to regular cow milk or ice cream
common_coldTotal	total number of episodes of common cold
egg	Age at first introduction, in days, of any egg
fruits_and_berries	Age at first intro to fruit and berries
rice_ricemilk	Age at first intro to rice milk
root_vegetables	Age at first intro to root vegetables
barley	Age at first introduction, in days, of any barley

Table A.2: Feature Description

Feature Name	**Description**
formula_any	Age at first intro to any formula
tot_duration	length of time supplements was taken
race_ethnicity	1:Hispanic, 2:White, 3:African American, 4:Other, 5: Unknown
sausage_hot_dogs	Age at first intro to any sausage or hot dogs
SadOrHappyFeelingDuringPreg	motherś feeling during pregnancy
impact	total number of child event that negatively affected parant or child
brst_fed	indicated child has stopped breast feeding
rice	Age at first intro to any rice
rice_milk	Age at first intro to rice or milk
illYesCount	number of different maternal illnees during pregnancy
Sex	gender
country	1:US, 2:Finland, 3:Germany, 4:Sweden
fibtime	Time to start formular or food
babysweightgrams	childś weight at birth in grams
HLA_Category	subjectś HLA category
cereals	Age at first intro to any cereals
maternal_age	motherś age at time of subjectś birth
fish_or_other_seafood	Age at first intro to fish or other seafood

Table A.3: Feature Description

Feature Name	**Description**
potatoes	Age at first intro to potatoes
time_to_brstfed_stop	time to stop breast feed
wheat	Age at first intro to any wheat
nongluten_cereals	Age at first intro to any nongluten creals
babyslengthcm	childś length in cm at birth
cow_milk_any	Age at first intro to any cow milk
gluten	Age at first intro to any gluten
ever_brstfed	If subject received breast milk
fdr	subjectś first degree relative status
milk_product	Age at first intro to any milk product
oat	Age at first intro to any oat
bmi_XX	Body mass index observed on month XX
inf_epi_C_XX	infection episode of disease C on month XX