Statistical Relational Learning to Predict Primary Myocardial Infarction from Electronic Health Records

Jeremy C. Weiss and **David Page**
University of Wisconsin-Madison
1300 University Ave, Madison, WI
{jcweiss,page}@biostat.wisc.edu

Sriraam Natarajan
Wake Forest University Baptist Medical Center
Medical Center Blvd, Winston-Salem, NC
snataraj@wakehealth.edu

Peggy L. Peissig
Marshfield Clinic Research Foundation
1000 North Oak Ave, Marshfield, WI
peissig@mcrf.mfldclin.edu

Catherine McCarty
Essentia Institute of Rural Health
502 E 2nd St, Duluth, MN
cmccarty@eirh.org

Abstract

Electronic health records (EHRs) are an emerging relational domain with large potential to improve clinical outcomes. We apply two statistical relational learning (SRL) algorithms to the task of predicting primary myocardial infarction. We show that one SRL algorithm, relational functional gradient boosting, outperforms propositional learners particularly in the medically-relevant high recall region. We observe that both SRL algorithms predict outcomes better than their propositional analogs and suggest how our methods can augment current epidemiological practices.

Introduction

One of the most studied pathways in medicine is the health trajectory leading to heart attacks, known clinically as myocardial infarctions (MIs). MIs are common and deadly, causing one in three deaths overall in the United States totaling 600,000 per year [Manson et al., 1992]. Because of its medical significance, MI has been studied in depth, mostly in the fields of epidemiology and biostatistics, yet rarely in machine learning. So far, it has been established that prediction of future MI is a challenging task. Risk stratification has been the predictive tool of choice [Group, 2002; Wilson et al., 1998], but these methods cannot reliably isolate the negative class; that is, everyone is still at risk. A much richer area of study is the identification of risk factors for MI. Common risk factors have been identified such as age, gender, blood pressure, low-density lipoprotein (LDL) cholesterol, diabetes, obesity, inactivity, alcohol and smoking. Studies have also identified less common risk factors as well as subgroups with particular risk profiles [Greenland et al., 2010; Antonopoulos, 2002].

The canonical method of study in this field is the identification or quantification of the risk attributable to a variable in isolation using: case-control studies, cohort studies, and randomized controlled trials. Case-control or cross-sectional studies identify odds ratios for the variable (or exposure) while controlling for confounders to estimate the relative risk. Cohort studies measure variables of interest at some early time point and follow the subjects to observe who succumbs to the disease. Randomized controlled trials are the gold standard for determining relative risks of single interventions on single outcomes. Each of these methods is highly focused, centered on the goal of providing the best risk assessment for one particular variable. One natural question to ask is: by using machine learning, can we conduct fewer studies by analyzing the effects of many variables instead?

Pt ID	Date	Diagnosis/Prescription/Procedure
207a3a567b745b50	2007.7	Lipitor
207a3a567b745b50	2010.8	Chest pain
207a3a567b745b50	2010.83	Angina pectoris
207a3a567b745b50	2011.2	Myocardial infarction

Pt ID	Date	Laboratory Test	Laboratory Value
207a3a567b745b50	2007.7	Cholesterol	High
207a3a567b745b50	2007.7	LDL	High
207a3a567b745b50	2008.7	LDL	Normal
207a3a567b745b50	2010.83	LDL	Normal

Pt ID	Gender	Date of Birth
207a3a567b745b50	Male	1962.34

Pt ID	Date	Vital Type	Vital Value
207a3a567b745b50	2007.7	BP	High
207a3a567b745b50	2007.7	BMI	Overweight
207a3a567b745b50	2008.7	BP	Normal
207a3a567b745b50	2010.83	BP	High

Figure 1: Can we use EHR data to augment clinical studies?

A different and crucial limitation of the longitudinal methods is that they make measurements at fixed points in time. In these studies, data is collected at the study onset t_0 to serve as the baseline variables, whose values are the ones used to determined risk. To illustrate this, consider the Skaraborg cohort study [Bg-Hansen et al., 2007] for the identification of acute MI mortality risk factors. The study

measured established risk factors for MI at t_0, and then the subjects participated in annual checkups to assess patient health and determine if an MI event had occurred. It is important to note that, in line with current practice, the subjects who did not possess risk factors at time t_0 but developed them at some later time were considered as not possessing them in the analysis. If we knew that these developments had occurred, say from an EHR, would it be possible to estimate the attributable risk more precisely? In the extreme, can we estimate the risk factors and make reliable predictions without the annual checkups and the baseline t_0 measurements?

More generally, can we bring a machine learning perspective to this task that provides new insights to the study of MI prediction and risk factor identification? The answer is yes, and we present here a glimpse of the potential machine learning has to bring to this field. We suggest that the emergence of the EHR as the new data source for population health analyses may be able to answer these clinical questions more efficiently, effectively adding another method of study to the standard three. For the prediction task, we emphasize the evaluation of methods on statistics that are clinically relevant, specifically on class separability (for risk stratification) and precision at high recalls (for use as a screening tool). Class separability, which can be directly assessed using ROC curves, is a well-established tool for risk stratification [Group, 2002]. Evaluating precision at high recalls assesses an algorithm's ability to predict while disallowing many false negatives, which is the critical component to a good screening tool. For predicting MI, a false negative means categorizing a patient as "low-risk" who goes on to have a heart attack, a costly outcome we wish to avoid. We also focus our methodology on algorithms with good interpretability, as this is critical for using the models for risk factor identification. In this work we survey a host of established machine learning algorithms for their performance on this task and select the most promising algorithm for further analysis. We attempt to answer some of these questions by providing an EHR-based framework for prediction and risk factor identification.

EHRs are an emerging data source of great potential use in disease prevention. An EHR effectively tracks the health trajectories of its patients through time for cohorts with stable populations (Figure 1). But as of yet they have been used primarily as a data warehouse for health queries, rather than as a source for population-level risk assessment and prevention. This trend is changing, however, as exemplified by the ongoing Heritage Health Prize contest, which uses medical claims data to predict future hospitalization [2011].

Analogously, we can use EHR data to predict future disease onset and produce personalized risk scores. Risk stratification models do exist, but they typically require additional medical intervention, e.g. running laboratory tests required to quantify risk. Thus, one advantage of risk modeling from EHRs is that many of the interventions required for standard risk stratification are rendered superfluous. While interventions provide up-to-date information and could improve risk stratification, a risk profile without them based on EHR data would be available regardless. As an example, the Framingham risk score (FRS) assessment of 10-year risk of coronary heart disease (CHD) requires two lipoprotein laboratory assays, blood pressure measurements, and basic demographic information [Antonopoulos, 2002]. The FRS is a well-established, robust test for evaluating CHD risk, but a similar risk could be established with information from the EHR, which might include overlapping clinical information, without the additional intervention. Furthermore, the ability to account for disease occurrences across time instead of the disease state at an initial time could help modify risk status. Finally, for less high-impact diseases than MI, the medical field has focused largely on identifying individual risk factors for disease. Relational models using EHRs could then easily produce aggregate risk models analogous to the FRS.

Accurate predictive models of MI or other major health events have many more potential applications. First, such models can be incorporated into the EHR to provide prompts to clinicians such as, "your patient is at high risk for an MI and is not currently on an aspirin regimen." Second, the models themselves can be inspected in order to identify surprising connections, such as a correlation between the outcome and the use of certain drugs, which might in turn provide important clinical insights. Third, these models can be used in research to identify potential subjects for research studies. For example, if we want to test a new therapy for its ability to prevent an event such as MI, it would be most instructive to test it in a population of high-risk subjects, which a predictive model can accurately identify.

The primary approach we use draws from relational probabilistic models, also known as *Statistical Relational Learning* (SRL) [Getoor and Taskar, 2007]. Their primary advantage is their ability to work with the structure and relations in data; that is, information about one object helps the learning algorithms to reach conclusions about other objects. Unfortunately, most SRL algorithms have difficulty scaling to large data sets. One efficient approach that yields good results from large data sets is the relational probability tree [Neville et al., 2003]. The performance increase observed moving from propositional decision trees to forests is also seen in the relational domain [Anderson and Pfahringer, 2009; Natarajan et al., 2010]. One method called functional gradient boosting (FGB) has achieved good performance in the propositional domain [Friedman, 2001]. We apply it to the relational domain for our task: the prediction and risk stratification of MI from EHRs.

EHR data presents significant challenges to current machine learning methodology. If we hope to augment traditional clinical study analyses, we must be able to effectively address these challenges. A few of them are: size, time-stamped data, relational data, and definition shifts over time. We use Relational Functional Gradient Boosting (RFGB) because it addresses all but the last challenge, which is difficult for any algorithm to capture. Notably, it is one of the few relational methods capable of learning from large data sets. Moreover, RFGB can efficiently incorporate time by introducing temporal predicates like *before*(A, B):-$A < B$. Also, unlike most other state-of-the-art SRL algorithms, RFGB allows us to learn structure and parameters simultaneously and grows the number of models as needed. Hence,

we apply RFGB [Natarajan et al., 2010] and relational probability trees (RPTs) [Neville et al., 2003] to the task of predicting primary myocardial infarction (MI). Our goal is to establish that, even for large scale domains such as EHRs, that relational methods, and in particular RFBG and RPTs, can scale and outperform propositional variants.

This paper makes a few key contributions: First, we address the challenging problem of predicting MI in real patients and identify ways in which machine learning can augment current metholodies in clinical studies. Second, we address this problem using recently-developed SRL techniques, adapt these algorithms to predicting MI and present the algorithms from the perspective of this task. Third, the task of MI prediction is introduced to the SRL community. To our knowledge, this is the first work to use SRL methods to predict MI in real patients. Fourth, we focus our analysis on interpretable RPT models, making it easy to discern the relationship between different risk factors and MI. Finally, our paper serves as a first step to bridge the gap between SRL techniques and important, real-world medical problems.

Learning Algorithms

Relational Probability Trees

RPTs [Neville et al., 2003] were introduced for capturing conditional distributions in relational domains. These trees upgrade decision trees to the relational setting and have been demonstrated to build significantly smaller trees than other conditional models and obtain comparable performance. We use a version of RPTs that employs the TILDE relational regression (RRT) learner [Blockeel and Raedt, 1998] where we learn a regression tree to predict positive examples (in this case, patients with MI) and turn the regression values in the leaves into probabilities by exponentiating the regression value and normalizing them. Hence, the leaves of the RPTs are still the probability that a person has an MI given the other attributes. The key advantage of TILDE is that it can use conjunctions of predicates in the inner nodes as against a single test by the traditional RPT learner. This modification has been shown to have better performance than RPTs by others [Natarajan et al., 2010; Anderson and Pfahringer, 2009]. In RRTs, the inner nodes (i.e., test nodes) are conjunctions of literals and each RRT can be viewed as defining several new feature combinations, one corresponding to each path from the root to a leaf. The resulting potential functions from all these different RRTs still have the form of a linear combination of features but the features can be quite complex [Gutmann and Kersting, 2006]. We use weighted variance as the criterion to split on in the inner nodes. We augment the RRT learner with aggregation functions such as *count, max, average* that are used in the standard SRL literature [Getoor and Taskar, 2007] thus making it possible to learn complex features for a given target. These aggregators are pre-specified and the thresholds of the aggregators are automatically learned from the data. Continuous features such as *cholesterol* level, *ldl*, *bmi*, etc. are discretized into bins based on domain knowledge.

Relational Functional Gradient Boosting

Assume that the training examples are of the form $(\mathbf{x}_i, y_i)$ for $i = 1, ..., N$ and $y_i \in \{0, 1\}$ where $y = MI$ and $\mathbf{x}$ represents the set of all observations about the current patient i. The goal is to fit a model $P(y|\mathbf{x}) \propto e^{\psi(y,\mathbf{x})}$. The standard method of supervised learning is based on gradient-descent where the learning algorithm starts with initial parameters θ_0 and computes the gradient of the likelihood function. A more general approach is to train the potential functions based on Friedman's gradient-tree boosting algorithm where the potential functions are represented by sums of regression trees that are grown stage-wise [Friedman, 2001]. More formally, functional gradient ascent starts with an initial potential ψ_0 and iteratively adds gradients Δ_i. Thus, after m iterations, the potential is given by $\psi_m = \psi_0 + \Delta_1 + ... + \Delta_m$. Here, Δ_m is the functional gradient at episode m and is

$$\Delta_m = \eta_m \times E_{x,y}[\partial / \partial \psi_{m-1} log\, P(y|x; \psi_{m-1})] \quad (1)$$

where η_m is the learning rate. Dietterich et al. [2004] suggested evaluating the gradient at every position in every training example and fitting a regression tree to these derived examples i.e., fit a regression tree h_m on the training examples $[(x_i, y_i), \Delta_m(y_i; x_i)]$. They point out that although the fitted function h_m is not exactly the same as the desired Δ_m, it will point in the same direction, assuming that there are enough training examples. So ascent in the direction of h_m will approximate the true functional gradient. The same idea has later been used to learn several relational models and policies [Natarajan et al., 2010; Sutton et al., 2000; Kersting and Driessens, 2008; Natarajan et al., 2011; Gutmann and Kersting, 2006].

Let us denote the MI as y and it is binary valued (i.e., occurrence of MI). Let us denote all the other variables measured over the different years as $\mathbf{x}$. Hence, we are interested in learning $P(y|\mathbf{x})$ where $P(y|\mathbf{x}) = e^{\psi(y;\mathbf{x})} / \sum_y e^{\psi(y;\mathbf{x})}$. Note that in the functional gradient presented in Equation 1, the expectation $E_{x,y}[..]$ cannot be computed as the joint distribution $P(\mathbf{x}, \mathbf{y})$ is unknown. Hence, RFGB treats the data as a surrogate for the joint distribution.

Instead of computing the functional gradients over the potential function, they are instead computed for each training example i given as $(\mathbf{x_i}, y_i)$. Now this set of local gradients form a set of training examples for the gradient at stage m. Recall that the main idea in the gradient-tree boosting is to fit a regression-tree on the training examples at each gradient step. In this work, we replace the propositional regression trees with relational regression trees [Gutmann and Kersting, 2006; Natarajan et al., 2010; Kersting and Driessens, 2008].

The functional gradient with respect to $\psi(y_i = 1; \mathbf{x_i})$ of the likelihood for each example $(\mathbf{x}_i, y_i)$ can be shown to be:

$$\frac{\partial \log P(y_i; \mathbf{x_i})}{\partial \psi(y_i = 1; \mathbf{x_i})} = I(y_i = 1; \mathbf{x_i}) - P(y_i = 1; \mathbf{x_i}),$$

where I is the indicator function that is 1 if $y_i = 1$ and 0 otherwise. The expression is very similar to the one derived in Dietterich et al. [2004]. The key idea in this work is to represent the distribution over MI of a patient as a set of RRTs

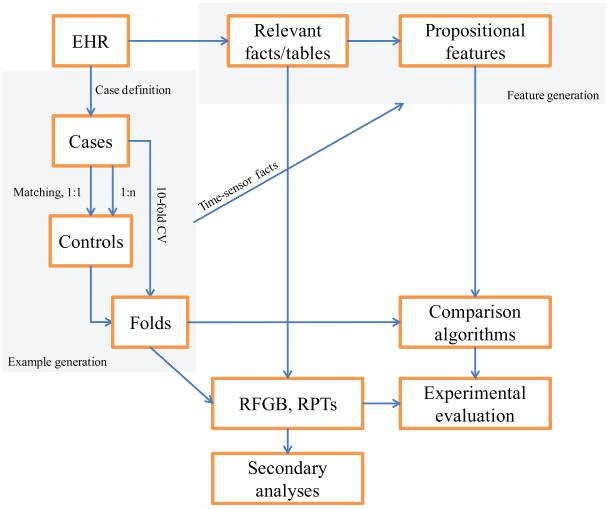

Figure 2: Flow chart depicting experimental setup

on the features. These trees are learned such that at each iteration the new set of RRTs aim to maximize the likelihood of the distributions with respect to ψ. Hence, when computing $P(MI(X)|\mathbf{f}(X))$ for a particular patient X, given the feature set $\mathbf{f}$, each branch in each tree is considered to determine the branches that are satisfied for that particular grounding (x) and their corresponding regression values are added to the potential ψ.

Experimental Methods

We analyzed de-identified EHR data on $18,386$ subjects enrolled in the Personalized Medicine Research Project (PMRP) at Marshfield Clinic [McCarty et al., 2005; 2008]. The PMRP cohort is one of the largest population-based biobanks in the United States and consists of individuals who are 18 years of age or older, who have consented to the study and provided DNA, plasma and serum samples along with access to their health information in the EHR. Most of the subjects in this cohort received most, if not all, of their medical care through the Marshfield Clinic integrated health care system.

Within the PMRP cohort, 1153 cases were selected using the first International Classification of Diseases 9th revision (ICD9) code of 410.0 through 410.1. Cases were excluded if the incident diagnosis indicated treatment for sequelae of MI or "MI with subsequent care". The age of the first case diagnosis was recorded and used to right-censor EHR data from both the case and the matching control one month prior to the case event. In other words, all facts linked to the case and the matched controls after the case age–one month prior to case diagnosis–were removed so that recent and future events could not be used in MI prediction.

To achieve a 1-1 ratio of cases to controls (i.e., positive and negative examples), cases were matched with controls based on the last age recorded in the EHR. For many matches, this corresponds to a case who is alive being matched to a control of the same age. For others it means matching someone who died from a heart attack to someone who died from other causes or was lost to follow-up. Matching on last reported age was chosen so that each subject would have both a similar age and similar presence in the EHR.

As CHD is the leading cause of mortality in the US, of which MI is a primary component, risk factors are well-studied [Antonopoulos, 2002; Greenland et al., 2010; Manson et al., 1992; Wilson et al., 1998], and those represented in the EHR were included in our experiments. We included major risk factors such as cholesterol levels (LDL in particular), gender, smoking status, and systolic blood pressure, as well as less common risk factors such as history of alcoholism and procedures for echocardiograms and valve replacements. Drugs known to have cardiac effects were included, notably the coxibs and tricyclic antidepressants. As EHR literals are coded in hierarchies, we chose to use the most specific level of information, which often split established risk factors into multiple subcategories. The risk factors were chosen a priori as opposed to employing algorithmic feature selection (e.g. the feature selection inherent in decision trees) to shrink the feature size from hundreds of thousands (excluding genetic data) to thousands for computational reasons and so that algorithms without inherent feature selection would perform comparably. The features chosen came from relational tables for diagnoses, medications, labs, procedures, vitals, and demographics.

Patient relations were extracted to temporally-defined features in the form of "patient ever had $x \in X$" or "patient had $x \in X$ within the last year". For laboratory values and vitals, both of which require an additional literal for the result of the test, the result was binned into established value categories (e.g. for blood pressure, we created five binary features by mapping the real value to {critically high, high, normal, low, and critically low}). This resulted in a total of 1,528 binary features.

The cases and controls were split into ten folds for cross-validation in a nine-fold train set to one-fold test set. Although we did choose a one-to-one ratio of cases to controls, in general this would not be the case, so we chose to assess the performance of the algorithms with the area under the ROC curve (AUC-ROC), accuracy, and by visualizing the results with a precision-recall plot. Also, precision at high recalls {0.95, 0.99, 0.995} were calculated to assess a model's usefulness as a screening tool. p-values were calculated comparing the RFGB model with the comparison methods using a two-sided paired t-test on the ten-fold test sets, testing for significant differences in accuracy and precision at a recall of 0.99.

The key question is whether the relational algorithms consistently produced better predictions than their corresponding propositional variant. Thus we compared RFGB models to boosted decision trees (AdaBoostM1 (Ada); default parameters) and RPTs with decision tree learners (J48; C=0.25, M=2). We also included other common models: Naive Bayes (NB; default parameters), Tree-Augmented Naive Bayes (TAN; SimpleEstimator), support vector machines (SVMs; linear kernel, C 1.0; radial basis function kernel, C 250007, G 0.01), and random forests (RF; 10 trees,

Table 1: Area under the ROC curve, accuracy and corresponding p-value(RFGB vs. all), precision at recalls (P@R), and p-value(RFGB vs. all, P@R=0.99). Bold indicates best performance.

	AUC-ROC	Accuracy	p	P@R=0.95	P@R=0.99	P@R=0.995	p(P@R=0.99)
Tree J48	0.744	0.716	4e-5	0.500	0.500	0.500	6e-7
Boosted Trees	0.807	0.753	1e-4	0.634	0.572	0.532	4e-4
Random Forests	0.826	0.785	4e-1	0.669	0.593	0.525	2e-3
NB	0.840	0.788	8e-1	0.680	0.513	0.500	1e-4
TAN	0.830	0.768	6e-3	0.662	0.518	0.500	2e-4
SVM (linear)	0.704	0.704	5e-6	–	–	–	–
SVM (rbf)	0.761	0.761	1e-2	–	–	–	–
RFGB	**0.845**	**0.791**	–	**0.688**	**0.655**	**0.625**	–
RPT	0.792	0.738	4e-6	0.622	0.595	0.578	4e-5

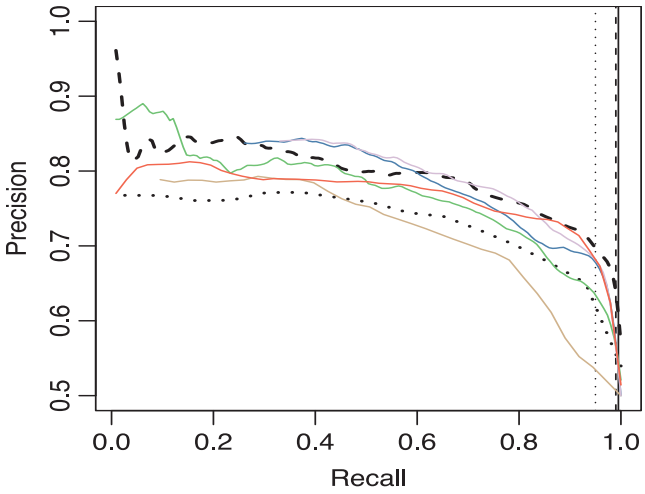

Figure 3: Precision-recall curves, with vertical lines denoting the recall thresholds {0.95, 0.99, 0.995}. RFGB (dashed) and RPT (dotted) are bolded. RFGB outperforms all other algorithms in the medically-relevant region (high recall). At recall=0.9, the ordering of algorithms (best to worst) is: RFGB, Random Forests, TAN, NB, RPT, Boosted Trees, J48.

default parameters). All propositional learners were run using Weka software [Hall et al., 2009]. In our secondary analysis, we varied both the experimental setup and the RFGB parameters to investigate the effect on their predictive ability. First, we altered the case-control ratio {1:1, 1:2, 1:3}, holding the number of cases fixed. Second, we altered the maximum number of clauses (for internal node splits) allowed per tree {3, 10 (default), 20, 30}. Third, we altered the maximum depth of the tree {1 (stump), 5}. Finally, we altered the number of trees {3, 10 (default), 20, 30}. We also compared the results among these analyses if they contained the same maximum number of parameters (e.g. 30 parameters: 3 trees $\times$ 10 clauses, 10 trees $\times$ 3 clauses).

Table 2: Secondary analyses: RFGB performance as case-control ratio (CC), number of clauses, trees and tree depth are modified. Default number of clauses = 10 and trees = 10

	AUC-ROC	Accuracy	P@R=0.99
CC 1:1;1:2;1:3	.84;.87;.88	.79;.80;.82	.66;.51;.43
Trees 3;20;30	.80;.85;.85	.74;.80;.80	.61;.67;.66
Clauses 3;20;30	.85;.85;.85	.79;.79;.79	.66;.66;.66
Tree depth 1;5	.85;.85	.79;.79	.66;.66

Results

The best cross-validated predictor of primary MI according to AUC-ROC was the RFGB model as shown in Table 1. RFGB outperformed the other tree learners, forest learners and SVMs. The RPT model did not score as well, ranking in the middle of the propositional learners. It is of note that the RFGB and RPT models significantly outperformed their direct propositional analogs (Boosted Tree and Tree models, respectively). The Bayesian model (NB; TAN) scores may be somewhat inflated because only features known to be CHD risk factors were specifically chosen for this analysis. They may be more prone to irrelevant feature noise as those models include all features into their final models.

The precision-recall curves for the algorithms are shown in Figure 3 (SVMs are omitted as their outputs do not admit a ranking over examples). Medically, the most important area is the region of high recall (i.e. sensitivity) because typically the cost of leaving a condition undiagnosed is high. In other words, the expected cost of a false positive is much smaller than a false negative because a false positive incurs the costs of additional interventions, while a false negative incurs costs of untreated human morbidity, and usually expensive, delayed treatments. Given that we cannot accept models with many false negatives (i.e. low recall), we look to the high recall region for the best performing algorithm, and RFGB gives the highest precision as shown in Table 1.

In our secondary analysis, when changing the case-control ratio we observed an increase in the AUC-ROC as well as the expected increase in accuracy and decrease in precision shown in Table 2. We suspect the improve-

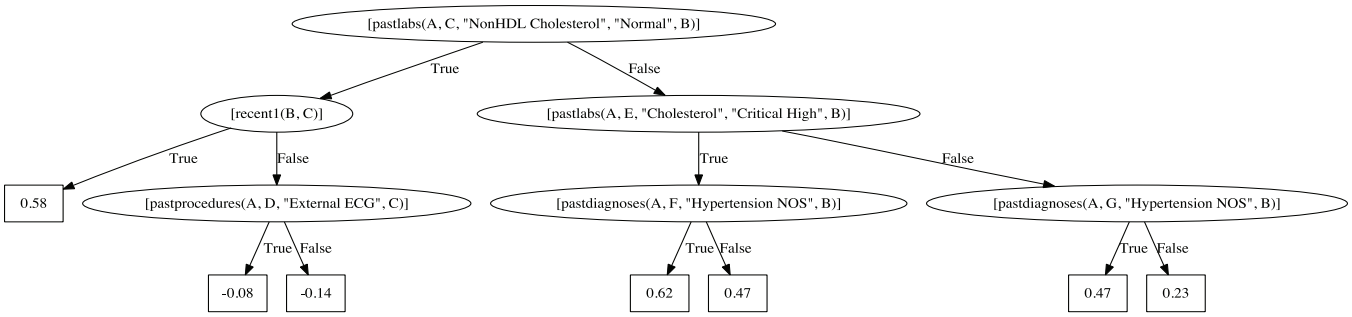

Figure 4: The first learned tree in the RFGB forest

ment in AUC-ROC may be attributed to the larger population size, as for example CC 1:3 has twice as many examples as CC 1:1. RFGB performance improved with increases with forest size, with the greatest gains coming between using three and ten trees, and no overfitting was observed using our largest fifty-tree forest (see our website: http://cs.wisc.edu/~jcweiss/iaai2012). Varying the number of clauses or tree depth made no visible difference in RFGB performance, at least when holding the number of trees fixed at ten. Per parameter, we found that increasing forest size improved prediction more than increasing individual tree sizes, as we see by comparing equal-parameter rows in Table 2.

Figure 4 shows an example tree produced in the RFGB forest. We can read this as follows. Given a patient A and their censor age B (i.e. for cases, one month before their first MI; for controls, the censor age of the corresponding case), if A had a normal non-HDL cholesterol measurement at time C, take the left branch, otherwise take the right branch. Assuming we took the left branch, if the measurement C was within one year of the censor age, take the left branch again. The leaf regression value is the best estimate of the residual of the probability of the covered examples given the model at that iteration. The whole RFGB forest is available at our website: http://cs.wisc.edu/~jcweiss/iaai2012.

Direct interpretation of the tree can lead to useful insights. In the example above, the tree indicates that a patient is more likely to have a future MI event if they have had a normal non-HDL cholesterol level reading in the last year compared to patients who have had normal cholesterol readings not in the last year. Now, since it is implausible that the measurement itself is causing MI, it could be considered a proxy for another "risk factor", which in this case could be physician concern, as frequent lipoprotein measurements may display a concern for atherosclerosis-related illness. The set of trees can also be converted into a list of weighted rules to make them more interpretable [Craven and Shavlik, 1996].

The density plot in Figure 5 shows the ability of RFGB and RPT models to separate the MI class from the controls. It is clear from the far left region of the RFGB graph that we can accurately identify a substantial fraction of controls with few cases by thresholding around 0.25, or more stringently at 0.05. This region captures an algorithm's utility as a screening tool, where we see that RFGB significantly outperforms the others.

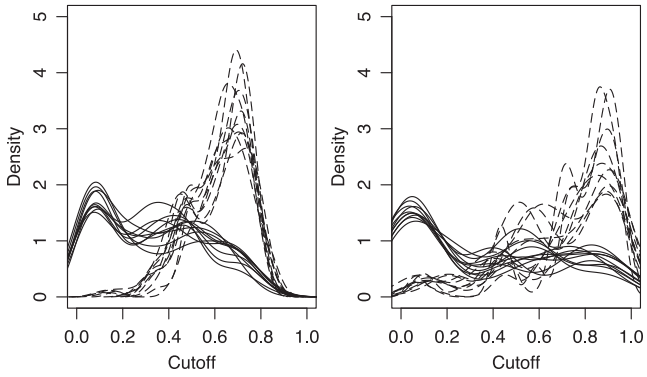

Figure 5: Density of cases (dashed) and controls (solid) by {RFGB (left), RPT (right)} prediction, one line per fold. Taking the integral from 0 to cutoff c for example at $c = 0.05$ and $c = 0.25$ show that RFGB identifies many controls at low-risk of developing MI.

Discussion and Conclusion

In this paper, we presented the challenging and high-impact problem of primary MI from an EHR database using a subset of known risk factors. We adapted two SRL algorithms in this prediction problem and compared them with standard machine learning techniques. We demonstrated that RFGB is as good as or better than propositional learners at the task of predicting primary MI from EHR data. Each relational learner does better than its corresponding propositional variant, and in the medically-relevant, high recall region of the precision-recall curve, RFGB outperforms all the other methods that were considered.

One additional layer of complexity not addressed in this experiment is the use of other relational information such as hierarchies. EHRs have hierarchies for diagnoses, drugs, and laboratory values, and it is important to be able to capture detail at each level. For example, characteristic disease progression pathways stem from infarctions of different heart walls, but at a high level, the presence of any MI leads to standard sequelae. Relational domains can easily incorporate this knowledge into hierarchical "is a" relations, whereas propositional learners must create new fea-

tures for every level. The challenge for relational tree-based learners is that the search algorithm is greedy; identifying high-level relations requires traversing several "is a" relationships first, and thus they might not be found in a greedy search. Expanding internal nodes to longer clauses has been implemented with some success [Natarajan et al., 2010; Anderson and Pfahringer, 2009], although this does have the effect of rapidly increasing the number of features to consider during branching. The use of SRL algorithms could also allow the use of relations like patient physicians and providers, which form complex relations less "patient-disease"-oriented but ones that still may be central to patient care. Questions regarding disease heritability could also be addressed through relational family-based analyses.

Given our initial success, we plan to extend our work by including more potential risk factors for learning (i.e., include all the measurements on all the patients). This will be challenging as the number and frequencies of the measurements will differ greatly across patients. In our current model, we used time as the last argument of our predicates. While there is a vast body of work in learning and reasoning with temporal models in propositional domains, the situation is not the same for relational models. We plan to investigate a principled approach to learning and reasoning with relational dynamic models that will allow physicians to monitor the cardiovascular risk levels of patients over time and develop personalized treatment plans. Finally, we plan to build a complete machine learning system for identifying risk factors across many diseases given the longitudinal data available in the EHR.

Acknowledgments

We acknowledge the gracious support by the grant 1UL1RR025011 from the Clinical and Translational Science Award (CTSA) program of the National Center for Research Resources, National Institutes of Health, the Wisconsin Genomics Initiative, NIGMs grant R01GM097618-01 and NLM grant R01LM011028-01, the Translational Science Institute (TSI) of Wake Forest Baptist Health, the Computation and Informatics in Biology and Medicine training grant 5T15LM007359, and the University of Wisconsin-Madison Medical Scientist Training Program.

References

Anderson, G., and Pfahringer, B. 2009. Relational random forests based on random relational rules. In *Interational Joint Conferences in Artificial Intelligence*.

Antonopoulos, S. 2002. Third report of the national cholesterol education program (ncep) expert panel on detection, evaluation, and treatment of high blood cholesterol in adults (adult treatment panel iii) final report. *Circulation* 106(3143):3421.

Blockeel, H., and Raedt, L. D. 1998. Top-down induction of first-order logical decision trees. *Artificial Intelligence* 101:285–297.

Bg-Hansen, E.; Larsson, C. A.; Gullberg, B.; Melander, A.; Bostrm, K.; Rstam, L.; and Lindblad, U. 2007. Predictors of acute myocardial infarction mortality in hypertensive patients treated in primary care. *Scandinavian Journal of Primary Health Care* 25(4):237–243.

Craven, M., and Shavlik, J. 1996. Extracting tree-structured representations of trained networks. In *Neural Information Processing Systems*, 24–30.

Dietterich, T.; Ashenfelter, A.; and Bulatov, Y. 2004. Training conditional random fields via gradient tree boosting. In *International Conference on Machine Learning*.

Friedman, J. H. 2001. Greedy function approximation: A gradient boosting machine. *Annals of Statistics* 1189–1232.

Getoor, L., and Taskar, B. 2007. *Introduction to Statistical Relational Learning*. MIT Press.

Greenland, P.; Alpert, J. S.; Beller, G. A.; Benjamin, E. J.; Budoff, M. J.; Fayad, Z. A.; Foster, E.; Hlatky, M.; Hodgson, J. M. B.; and Kushner, F. G. 2010. 2010 accf/aha guideline for assessment of cardiovascular risk in asymptomatic adults. *Journal of the American College of Cardiology* j. jacc. 2010.09. 001v1.

Group, D. P. C. 2002. Prediction of mortality from coronary heart disease among diverse populations: is there a common predictive function? *Heart* 88:222–228.

Gutmann, B., and Kersting, K. 2006. TildeCRF: Conditional Random Fields for Logical sequences. In *European Conference on Machine Learning*.

Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; and Witten, I. H. 2009. The weka data mining software: an update. *Special Interest Group on Knowledge Discovery and Data Mining Explorations Newsletter* 11(1):10–18.

Heritage Provider Network, I. 2011. Heritage health prize.

Kersting, K., and Driessens, K. 2008. Non–parametric policy gradients: A unified treatment of propositional and relational domains. In *International Conference on Machine Learning*.

Manson, J. A. E.; Tosteson, H.; Ridker, P. M.; Satterfield, S.; Hebert, P.; O'Connor, G. T.; Buring, J. E.; and Hennekens, C. H. 1992. The primary prevention of myocardial infarction. *New England Journal of Medicine* 326(21):1406–1416.

McCarty, C. A.; Wilke, R. A.; Giampietro, P. F.; Wesbrook, S. D.; and Caldwell, M. D. 2005. Marshfield clinic personalized medicine research project (pmrp): design, methods and recruitment for a large population-based biobank. *Personalized Medicine* 2(1):49–79.

McCarty, C. A.; Peissig, P.; Caldwell, M. D.; and Wilke, R. A. 2008. The marshfield clinic personalized medicine research project: 2008 scientific update and lessons learned in the first 6 years. *Personalized Medicine* 5(5):529–542.

Natarajan, S.; Khot, T.; Kersting, K.; Guttmann, B.; and Shavlik, J. 2010. Boosting Relational Dependency networks. In *Inductive Logic Programming*.

Natarajan, S.; Joshi, S.; Tadepalli, P.; Kristian, K.; and Shavlik, J. 2011. Imitation learning in relational domains: A functional-gradient boosting approach. In *Interational Joint Conferences on Artificial Intelligence*.

Neville, J.; Jensen, D.; Friedland, L.; and Hay, M. 2003. Learning Relational Probability trees. In *Knowledge Discovery and Data Mining*.

Sutton, R.; McAllester, D.; Singh, S.; and Mansour, Y. 2000. Policy gradient methods for reinforcement learning with function approximation. In *Neural Information Processing Systems*.

Wilson, P. W. F.; D'Agostino, R. B.; Levy, D.; Belanger, A. M.; Silbershatz, H.; and Kannel, W. B. 1998. Prediction of coronary heart disease using risk factor categories. *Circulation* 97(18):1837.