Survival Association Rule Mining Towards Type 2 Diabetes Risk Assessment

Gyorgy J. Simon, PhD[1], John Schrom[1], M. Regina Castro, MD[2],

Peter W. Li, PhD[2], Pedro J. Caraballo, MD[2]

[1]University of Minnesota, Minneapolis, MN; [2]Mayo Clinic, Rochester, Minnesota

Abstract

Type-2 Diabetes Mellitus is a growing epidemic that often leads to severe complications. Effective preventive measures exist and identifying patients at high risk of diabetes is a major health-care need.

The use of association rule mining (ARM) is advantageous, as it was specifically developed to identify associations between risk factors in an interpretable form. Unfortunately, traditional ARM is not directly applicable to survival outcomes and it lacks the ability to compensate for confounders and to incorporate dosage effects. In this work, we propose Survival Association Rule (SAR) Mining, which addresses these shortcomings.

We demonstrate on a real diabetes data set that SARs are naturally more interpretable than the traditional association rules, and predictive models built on top of these rules are very competitive relative to state of the art survival models and substantially outperform the most widely used diabetes index, the Framingham score.

Introduction

Diabetes mellitus is a growing epidemic that affects 25.8 million people in the United States (8% of the population)[4]. Diabetes leads to significant medical complications including ischemic heart disease, stroke, nephropathy, retinopathy, neuropathy and peripheral vascular disease. Appropriate management of patients at risk with lifestyle changes and/or medications can decrease the risk of developing diabetes by 30% to 60% [6,12]. Therefore, early identification of patients at risk of developing diabetes is therefore a major healthcare need. In response to this pressing need, numerous diabetes risk indices have been developed and some of them, most notably the Framingham score[13], have gained acceptance in clinical practice.

Existing diabetes indices largely assume that diabetes is independent of other diseases. As diabetes is part of the metabolic syndrome, it is particularly important to consider the possibility of interactions between various risk factors, many of which are also indicators of other diseases in the metabolic syndrome. Except for the most recent methods[3, 8, 10], no diabetes index takes the interactions between the risk factors into account[5].

Association rule mining[1] (ARM) is a technique that is specifically aimed at discovering interactions (more precisely, associations). In association rule mining, we first extract **association patterns**, which are co-occurring binary risk factors. For example, we may discover the pattern that hypertension (high blood pressure) and hyperlipidemia (high cholesterol) frequently co-occur in patients. The frequent co-occurrence of these two conditions may indicate that they are associated with each other. We have formal tests to determine whether a presumed association is significant or a mere coincidence[11]. If the pattern is predictive of an outcome of interest—diabetes in our case,—we can convert the pattern into an association rule. An **association rule** is an implication, where a pattern of co-occurring conditions implies increased risk of diabetes. Continuing with our example, we may find that among patients presenting with hypertension and hyperlipidemia, 13.3% have diabetes, which is 1.47 times higher than in the general population of our study, making this pattern predictive of diabetes. Association rule mining is rapidly gaining popularity in health informatics due to the ease of interpretation, the ability to discover potentially interesting associations among risk factors, and since the results are rules, they are amenable to implementation in a clinical decision support system. The above three recent metabolic syndrome studies[3, 8,10] all used association rule mining.

Association rule mining, in its current form, is not applicable to survival outcomes in a straightforward fashion. Association rules are often used for quantifying the risk that the constituent risk factors confer on the patient subpopulation using the simple method we presented earlier. Unfortunately, this approach is incorrect, as it fails to account for age as a risk factor: the hypertensive, hyperlipidemic subpopulation is older than the population without these conditions. A third shortcoming of the traditional association rule mining paradigm lies in its inability to capture dosage effects. Many risk factors in diabetes can have two effects: a dosage effect, where a unit increase in a

measurement is accompanied by a proportional increase (or decrease) in the risk; and a threshold effect, which arises when a measurement exceeding a threshold causes disproportionate increase in the risk. Figure 1 illustrates the effect of systolic blood pressure (SBP) on the risk of DM. The solid line depicts the smoothed risk of DM as a function of SBP. When SBP < 140 mmHg (the ADA[2] recommended cutoff), the average risk of DM is approximately 9%, which coincides with the risk in the general population. When SBP exceeds 140 mmHg, the risk is substantially higher on average. The dashed line in the Figure represents this threshold effect and it is emblematic of association rule mining's view of SBP. It is clear from the Figure, that besides the threshold effect, a dosage effect also exists and ignoring this dosage effect leads to information loss. Association rule mining in its current form operates on the threshold effects and possesses no facility to incorporate dosage effects.

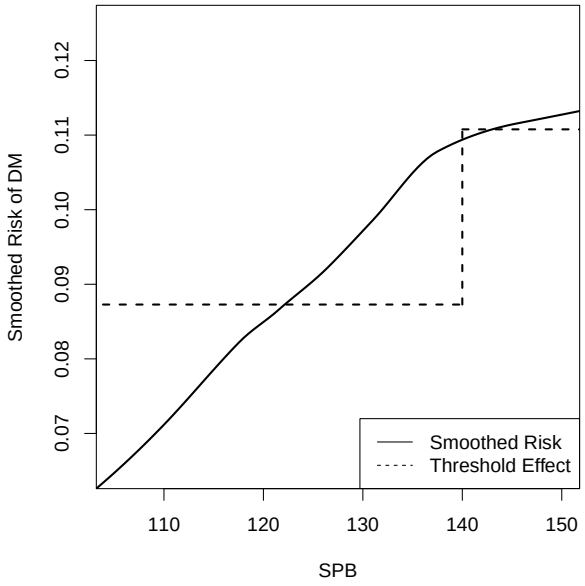

Figure 1. Illustration of the threshold effect. Risk of DM versus systolic blood pressure (SBP) is plotted for patients with SBP between 100 and 150 mmHg (90% of the population). When SBP is less than 140, the risk of DM on average is the same as the risk in the general population. Patients with SPB > 140, have a substantially higher risk on average.

In this work, we propose Survival Association Rule Mining, which extends the traditional association rule mining paradigm to address the above shortcomings. The key idea is transform the non-parametric and model-free association rule mining into a semi-parametric modeling paradigm. The centerpiece of this paradigm is the **survival association rule** (SAR), which is an association pattern wrapped into a survival model. The association pattern within the SAR captures the potential interaction among multiple binary risk factors (threshold effects) in a non-parametric fashion. The survival model around the association pattern links the predictors (covariates and the association rule) to the survival outcome, and provides a parametric ``interface'' to the rule that we can exploit towards adjustments for confounders and towards the incorporation of dosage effects. Thus a SAR preserves the flexibility of traditional association rule mining while it offers the advantages (e.g. adjustment for factors) that parametric models offer.

We applied the proposed methodology to a real clinical data set collected at Mayo Clinic. We show that the proposed approach identified clinically relevant association rules, and estimated the risk associated with the risk factors in the rules more correctly than traditional association rules in a manner that makes interpretation even simpler. We have also built predictive models for individual patients and we show that the resulting model performs as well as the state-of-the-art survival models and that the model-based association rules far outperformed the most widely adopted diabetes risk score, the Framingham score.

Background

Association rule mining

Consider a set of binary variables, called **items**, including abnormal laboratory results, history of diseases and whether a particular medication was prescribed. Let δ_j denote an item indicating whether patient *j* had developed diabetes.

An **itemset** (or **association pattern**) is simply a set of items. An itemset **covers** (or applies to) a patient, if the patient presents all the conditions in the itemset. The **support** of an itemset is the number of patients it applies to. The **coverage vector** X_I of an itemset *I* is a vector with its *j* element signifying whether *I* applies to the patient: it takes the value 1 if *I* covers patient *j*; 0 otherwise.

A **survival association rule** is an implication defined by an itemset *I*, stating that patients who suffer from the conditions in *I* have a significantly higher risk of diabetes than the average patient in our population. The increase in the risk for this subpopulation is quantified by the **relative risk**. Let *E* denote the expected and *O* the observed number of diabetes events (*O* takes the value 0 or 1) for each patient. The relative risk is defined as *RR=O/E*: patients who present the conditions in *I* faces *RR* times higher risk (number of diabetes events) than those who lack at least one condition.

Survival Models

Survival models are statistical models for event-based data with known time to events. Let t_j denote the follow-up time for patient *j*. The **follow-up time** is the time from the beginning of study until the patient progresses to diabetes or until he is no longer followed. Let δ_j denote whether patient *j* developed diabetes before (or exactly at) time t_j. The **hazard** $\lambda_j(t)$ is the instantaneous probability that patient *j* progresses to diabetes exactly at time *t*.

Cox Proportional Hazard Models are survival models that estimate the hazard $\lambda_j(t)$ for patient *j* at time *t* based on a covariate matrix *Z* and a baseline hazard $\lambda_0(t)$ that is common to all patients. The hazard is modeled as

$$\lambda_j(t) = \lambda_0(t)\exp\left(Z_j\beta\right),$$

where β is a coefficient vector to be estimated and the baseline hazard $\lambda_0(t)$ is unspecified. The quantity

$$r_j = Z_j\beta$$

is called the **risk**. For a patient *j*, the expected number of events (progression to diabetes) can be estimated based on his risk r_j as

$$\Lambda_j(t) = \sum_{k:t_k \le t_j} \frac{\delta_k \exp(r_j)}{\sum_{t_i \ge t_k} \exp(r_i)}.$$

The difference between the observed O_j and estimated number Λ_j of events (at the end of study)

$$M_j = O_j - \Lambda_j$$

is the **Martingale residual**.

The coefficient vector β is estimated through maximizing the partial likelihood. The **partial likelihood** of the data is defined as

$$PL(r_j) = \prod \left[\frac{\exp(r_j)}{\sum_{k:t_k \ge t_j} \exp(r_k)} \right]^{\delta_j},$$

where *k* is iterating through the patients who are at risk at time t_j. Once a patient progresses to diabetes, he is no longer at risk and he is no longer followed as far as this model is concerned.

Cox Proportional Hazards Models are fit by maximizing *PL(r)*, or equivalently, by minimizing the **negative log likelihood** ℓ of the data

$$\ell(r) = -\log PL(r).$$

The minimization can be carried out using Fisher Scoring for all variables at once, or stage-wise through gradient descent optimization, where in each iteration (stage), a new predictor is added. In the following section, we review the latter alternative in detail.

Gradient Boosting for Cox Proportional Hazards Models

Gradient boosting[7,9] minimizes ℓ iteratively through gradient descent optimization. Given a model in the *i*th iteration, it is extended with a fitted model $m^{(i)}$ that minimizes ℓ

$$r^{(i+1)} = r^{(i)} - \gamma m^{(i)} \approx r^{(i)} - \gamma \frac{\partial \ell(r)}{\partial \gamma}.$$

The fitted model $m^{(i)}$ minimizes ℓ when it is most aligned with the negative gradient of the negative log likelihood function evaluated at $r^{(i)}$, which is the Martingale residual defined earlier. In our application, $m^{(i)}$ will be a linear combination of covariates or a single itemset (represented by its coverage vector).

Methods

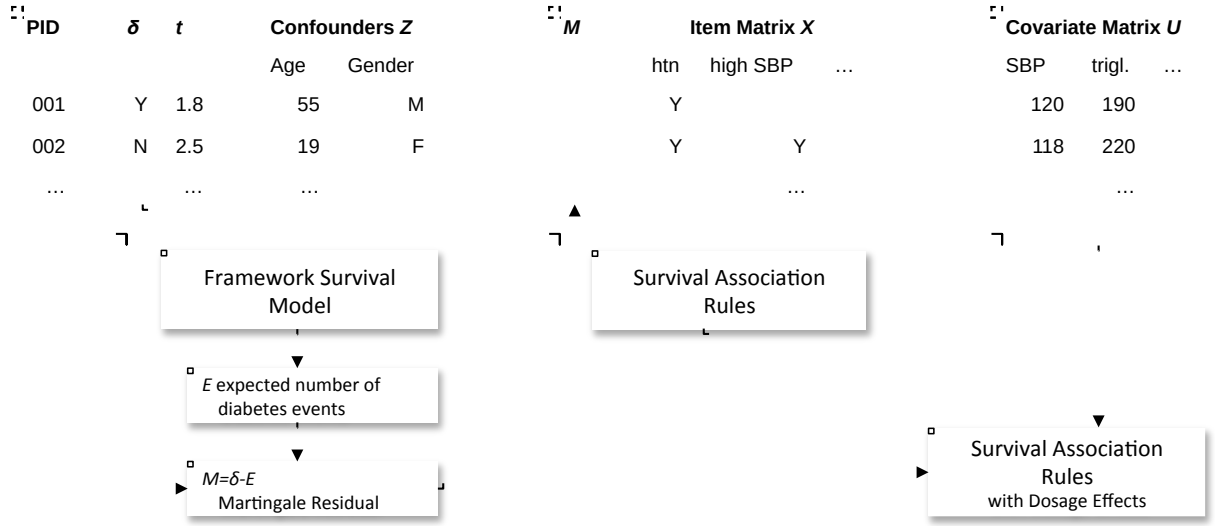

Figure 2. Overview of the methodology

We start the description of our methodology through presenting an overview in Figure 2. Let *Z* denote a covariate matrix of confounders (e.g. age and sex), *U* a covariate matrix of vitals and laboratory results and *X* an item matrix, contains the items representing medication prescriptions, abnormal laboratory results and the presence of co-morbid diagnosis codes. Note that *U* contains the continuous version and *X* the binary version of the vitals and the laboratory measurements. Let *t* denote the follow-up time, which is the time to developing diabetes or last follow up, and δ denote the diabetes outcome at last follow-up. The *j*th row of the matrices Z_j , X_j , U_j and the *j*th element of t_j and δ_j correspond to the same patient for all *j*.

We first build the **framework model** S_F, which provides a linkage between the predictors and the survival outcome (*t* and δ) and it also allows us to correct for the confounders in *Z*

$$S_F : Surv(\delta, t) = Z\beta.$$

We use the framework model to calculate the expected number *E* of diabetes events for each patient. By comparing the expected number of events to the observed events (δ), we arrive at the martingale residual *M*. The martingale

residual is the excess risk of diabetes that cannot be explained by follow-up time or by the confounders, but it can be at least partially explained by the covariates (binary or continuous).

With the set of association patterns in hand, we construct the survival association rules (SARs). A SAR is an extension of the framework model by a single association pattern (or itemset) I

$$S_A : Surv(\delta, t) = Z\beta - X_I\gamma ,$$

where X_I is the coverage vector of I and γ is its coefficient. Note that there is a single framework model, but there are multiple SARs; one SAR for each association pattern.

It is well known that a combinatorially large number of association patterns can be discovered from a database and many of these patterns may not explain the martingale residual significantly. Constructing a SAR for each discovered pattern is wasteful and computationally infeasible. We only construct a SAR from those patterns that significantly improve the fit of the framework model.

Since the survival association rule S_A is a model fully nested inside the framework model S_F, the significance of I can be assessed through a likelihood ratio test. The test statistic $2\log(\ell_F - \ell_A)$ follows Chi square distribution with 1 degree of freedom. The likelihood ratio test requires that S_A be fit exactly.

Fitting a survival model exactly is computationally expensive. In the following section of this work, we develop a metric, *gain*, based on gradient boosting that is almost perfectly correlated with the above likelihood ratio test statistic, does not require a survival model to be fit exactly and can be computed in a single scan of the coverage vector. *Gain* allows us to order the association patterns based on their significance and not fit a survival model exactly for the vast majority of the association patterns that would be found non-significant by the likelihood ratio test.

Once the survival association rules are computed, we can incorporate dosage effects in a relative straightforward manner. Suppose the survival association rule includes items that are dichotomized version of quantitative measures that we also have access to in U. Let U_I denote the columns of U that correspond to such items. We can incorporate dosage effects by extending the survival association rule by U_I with coefficients α

$$S_D : Surv(\delta, t) = Z\beta - X_I\gamma - U_I\alpha .$$

Deriving the gain metric

In this section, we derive *gain*, the metric that allows us to identify significant association patterns without having to fit a survival model to all discovered association patterns.

Given the fitted framework model S_F, with coefficient vector β, let r_F denote the predicted risk $r_F=Z\beta$. Suppose we are also given an association pattern I with coverage vector X_I. We can obtain the risk vector r_A as $r_A = r_F - \gamma X_I$, where γ is a scalar to be estimated. The typical method of finding γ is line-search to minimize $\ell(r_A)$. An alternative method can be derived from the Taylor expansion of ℓ

$$\ell(r - X_I\gamma) = \ell(r) - \gamma^T \frac{\partial \ell(r)}{\partial r} + \frac{1}{2}\gamma^T \frac{\partial^2 \ell(r)}{\partial r^2}\gamma = \ell(r) - \gamma^T X_I \frac{\partial \ell(r)}{\partial \gamma} + \frac{1}{2}\gamma^T X_I^T \frac{\partial^2 \ell(r)}{\partial \gamma^2} X_I\gamma .$$

By setting $r=r_A$, $\gamma=0$ and by simplifying the Hessian $\partial^2\ell(r)/\partial\gamma^2$ to the identity matrix, we can rewrite the Taylor expansion as

$$\ell(r_A - X_I\gamma) = \left(\frac{\partial \ell(r_A)}{\partial \gamma} - X_I\gamma\right)^T \left(\frac{\partial \ell(r_A)}{\partial \gamma} - X_I\gamma\right) = \left(M - X_I\gamma\right)^T \left(M - X_I\gamma\right),$$

which represents the gradient descent optimization as a least squared problem. The vector M denotes the vector of martingale residuals. The least squared representation allows us to solve for γ analytically, yielding the solution

$$\gamma = \left(X_I^T X_I\right)^{-1} X_I^T \frac{\partial \ell(r)}{\partial \gamma} = \left(X_I^T X_I\right)^{-1} X_I^T M \ .$$

We define **gain** as the reduction in the sum squared error due to adding I

$$G(I) = \left(X_I^T M\right)^T \left(X_I^T X_I\right)^{-1} X_I^T M \ .$$

Gain is almost perfectly correlated with the likelihood ratio test statistic yet it can be computed in a single scan of the coverage vector. While the gain cannot be used to assess whether an itemset improves the fit significantly (the p-value calculated based on $G(I)$ would be a poor estimate), it can be used to rank the itemsets almost perfectly based on their significance.

Strategy for Discovering Association Rules

The high correlation between the gain $G(I)$ and the likelihood ratio test statistics suggests the following strategy to discover *significant* association patterns effectively. The high correlation implies that if for a pair of association patterns, one has higher gain, it also has a higher likelihood ratio statistic and consequently, higher significance (lower p-value). We can also find the smallest gain g, such that the corresponding association patterns still significantly improves the fit of the framework model. Then for all patterns that have $G(I)<g$, the pattern is not significant. This gain g can be found through binary search, which requires fitting $\log_2 R$ survival models, where R is the number of discovered association patterns. This means that we need to fit survival models for at most $\log_2 R$ insignificant association patterns.

Making Predictions for Individual Patients

Although the main purpose of (survival) association rule mining is to discover interconnections between various risk factors that affect a patient's progression to diabetes, it is often useful to use the survival association rules as a diabetes index, where we need to predict the risk of progression for an individual patient. Since the survival association rules are survival models, they can readily predict the risk for each patient that they apply to. The only difficulty is that multiple rules may apply to the same patient. In that case predictions can be made using the most specific rule, that is, the rule that includes the highest number of conditions that the patient presents. In the case of ties (namely, when multiple applicable rules have the same number of conditions), we can take the average of the risks predicted by these rules.

Results

In this section, we demonstrate the above concepts on a clinical data set collected for a study at Mayo Clinic between 2005 and 2010. We will show that the survival association rules are more interpretable than the traditional association rules when used for assessing the risk of a patient subpopulation; and we also show that survival association rules are as predictive as survival models built on all predictors and substantially more predictive than the Framingham score.

Table 1. Descriptive Statistics for some of the continuous measures

Measure	Explanation	Descriptive		Abnormal		Missing	
		mean	sd	count	%	count	%
age		52.25	16.62	0	0.00	0	0.00
sbp	Systolic blood pressure	128.23	15.86	4573	0.21	1216	0.06
dbp	Diastolic blood pressure	77.16	9.13	1980	0.09	1216	0.06
tchol	Total cholesterol	199.44	30.95	9465	0.43	2919	0.13
hdl	High-density lipoprotein	53.65	13.65	5003	0.23	3004	0.14
ldl	Low-density lipoprotein	117.21	26.79	13856	0.63	3063	0.14
bmi	Body mass index	28.59	5.23	5941	0.27	4508	0.21
trigl	Triglycerides	139.38	67.48	6518	0.30	2972	0.14

The study is comprised of 21,981 patients. These are pre-diabetic patients who lived in Olmstead Co, MN in 2005. We established their pre-diabetic status by retrospectively collecting their fasting glucose measurement for the period of 1999/01/01 through 2004/12/31. These patients had at least one glucose measurement between 101 and

125 mg/dl and no measurements in excess of 126 mg/dl during this period. Patient with an established diabetes diagnosis during this period were excluded.

For our cohort, we collected demographic information (age, gender), vitals and laboratory results, diagnosis codes related to the metabolic syndrome and prescriptions. The most important variables are described in Tables 1 and 2. For association rule mining, the continuous variables were dichotomized at thresholds recommended by the ADA guidelines[2].

Table 2. Definition, abbreviation and prevalence of the most important binary variables

Predictor	Abbr.	%
demographic		
gender male	genderM	0.49
diagnoses		
hypertension	htn	0.31
hyperlipidemia	hyperlip	0.37
obese	obese	0.20
ischemic heart disease	ihd	0.10
peripheral vascular disease	pvd	0.02
medications		
ACE/ARB	acearb	0.13
beta blocker	bb	0.17
Ca channel blocker	ccb	0.07
diuretic	diuret	0.13
fibrate	fibra	0.02
statin	statin	0.18
aspirin	aspirin	0.29

Table 3. Ten traditional association rules with the highest relative risk

RR	sup	supD	itemset
2.35	293	62	hyperlip trigl fibra
2.29	301	62	trigl fibra
2.27	449	92	htn hyperlip bmi trigl aspirin
2.17	686	134	hyperlip bmi trigl aspirin
2.15	563	109	htn bmi trigl aspirin
2.13	496	95	htn hyperlip bmi trigl statin
2.10	502	95	htn bmi trigl statin
2.09	371	70	hyperlip fibra
2.04	669	123	htn obese hyperlip trigl
2.03	907	166	htn hyperlip bmi trigl

Results from Traditional Association Rule Mining

First, we present results obtained from traditional association rule mining. We discovered 2,054 association patterns. We filtered them based on predictive significance[11] and component independence[11]. A pattern is predictive significant if the conditional probability of diabetes given the itemset is statistically significantly different from the prior probability of diabetes; and an itemset $I = AB$ is component independent if it can be divided into sub-itemsets

(components) *A* and *B* that are statistically independent. Rules that are not predictive are clinically irrelevant and rules that are component independent are superfluous. After the filtering, we had 156 rules remaining. We refer to these rules as the **significant (traditional) association rules**.

Table 3 depicts the ten significant traditional association rules with the highest relative risk. The interpretation of (say) the first rule appears straightforward: a patient who is hyperlipidemic, has abnormally high triglicerides and takes fibrates faces on average 2.35 times higher risk of progression to diabetes than the average patient in our cohort. Unfortunately, this statement is true only if the patients we apply this rule to have the *same age and gender* distribution as the above subpopulation. This condition, which is often ignored, makes the application of the rule less practical.

Comparison of Traditional Association Rules and Survival Association Rules

We also extracted survival association rules from the data set. All 156 rules were found to significantly improve the fit of the framework model (at Bonferroni-adjusted .05 significance level).

The interpretation of the survival association rules is as follows. Hyperlipidemic patients with high triglyceride levels taking fibrates face a relative risk of 2.31, i.e. they have a 2.31 times higher risk of progression to diabetes than a patient with the same age and gender who does not have at least one of the above conditions. This statement is better suited for clinical application than the one for the traditional association rule.

In the above example, the difference in the relative risk estimate between the traditional and survival association rules is minimal (2.35 vs 2.31). The are not always minimal; in some cases they can become substantial. In Figure 3, we present a comparison: each point in the figure corresponds to one of the 156 survival association rules with the horizontal axis depicting the relative risks estimated by the traditional association rules and the vertical axis by the survival association rules.

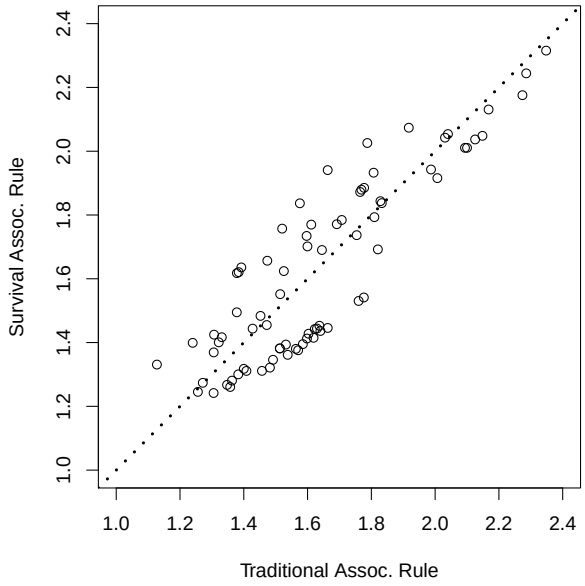

Figure 3. Relative risk estimates by traditional and survival association rules

In general the relative risk estimates are consistent between the two types of association rules: if survival association rule mining estimates a pattern to have high relative risk, traditional association rule mining also estimates it to have high relative risk. There are some notable exceptions, however. For example, the rule {htn, hyperlip, sbp, acearb, statin} carries a relative risk of 1.78 when estimated by traditional association rules and 1.54 when estimated by survival association rules. The former estimate corresponds to the 83rd percentile of the relative risks (only 27% of the traditional association rules predict a higher relative risk), while the latter estimate corresponds to the 49th percentile (more than half of the rules predict a higher relative risk). The discrepancy stems from the substantial age difference between the affected subpopulation (mean age is 69.8) and the unaffected subpopulation (mean age is 51.8). The relative risk estimate by traditional association rule mining for this subpopulation is misleading.

Evaluating the Predictive Capability of Survival Association Rules

Although our primary goal was to identify combinations of risk factors that confer a significantly elevated risk of progression to diabetes onto the patients, in this section, we present results in which we used the survival association rules as a diabetes index. The index score is the predicted risk of diabetes.

To assess the performance of the rules as a diabetes index, we performed 100 replications of the following procedure. We discovered the association patterns and fit the survival association rules on 80% of the patients (training set) and made predictions for the remaining 20%. A Lasso-penalized survival model was also constructed. Lasso-penalized models require that we tune a penalty parameter, to which end, we used 20% percent of the training set as a validation set. The evaluation metric is the **concordance** of the prediction with the true outcome. Concordance is the probability that for a pair of patients, where one progressed to diabetes and one did not, the patient who progressed has higher predicted risk.

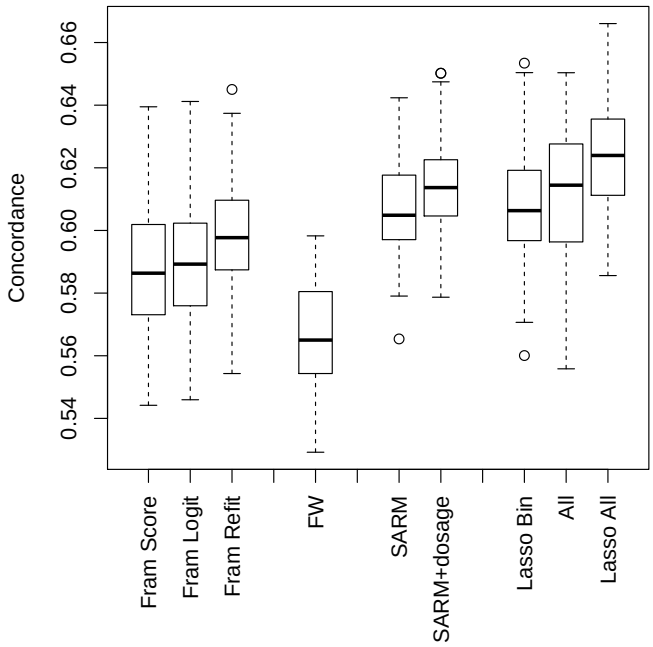

Figure 4. Comparing the concordance of the prediction by a number of risk models with the true outcome in 100 replications. The models used are Framingham Score, the logistic version of the Framingham score with the original coefficients (Fram Logit), the logistic version of the Framingham Score with coefficients re-fit to our data (Fram Refit), framework model (FW), Surivival association rules without dosage effect (SARM), and SAR with dosage effects (SARM+doage), Lasso-penalized survival models on all binary predictors (Lasso Bin), unpenalized survival model on all variables (All), and Lasso-penalized survival model on all variables (Lasso All)

In Figure 4, we present the performance of a number of algorithms and indices. The first three are the Framingham score and two of its variants: Fram Logit is the logistic regression form of the Framingham score with the original coefficients and Fram Refit is the logistic regression form with coefficients fit to our data. `FW' is the framework model, SARM and `SARM+dosage' correspond to survival association rules with and without dosage effects, respectively. `Lasso Bin' and `Lasso All' are Lasso-penalized survival models built on all binary predictors and all predictors (binary as well as quantitative). `All' is an unpenalized survival model that uses all predictors.

We can make the following observations. First, by comparing the framework model 'FW' with 'SARM', we can see that adding the association patterns to the framework model substantially improves the predictive performance. Comparing `SARM' with `SARM+dosage' demonstrates that incorporating dosage effects into the survival association rules substantially increased the models' predictive capability. Third, despite the very simple strategy we applied to make predictions for the individual patients, `SARM' performed as well as the state-of-the-art Lasso model `Lasso Bin'. Fourth, comparing `SARM+dosage' to `All' and `Lasso All' is possible but not entirely fair, as

`All' and `Lasso All' are more flexible models than `SARM+dosage': `SARM+dosage' only compensates for the dosage effect of risk factors that are present in the association pattern, while the `All' models can incorporate all dosage effects. Despite this difference, `SARM+dosage' performed identically to the unpenalized `All' model but performed worse than the penalized `Lasso All' model. This result cautions us that fitting dosage effects to small sub-population can lead to model overfitting and in the future, we will consider penalized survival association rules. Lastly, survival association rules, with or without dosage effects alike, substantially outperformed all three variants of the Framingham score that we considered.

Summary and Conclusion

Association rule mining is rapidly becoming a popular technique to analyze the interconnections between diseases and risk factors in a relatively interpretable form. In this work, we have presented an extension to the traditional association rule mining paradigm, which allows for (i) handling survival outcomes, (ii) making adjustment for confounders and (iii) incorporating dosage effects. We have shown that due to the adjustments, our rules are more interpretable and more suitable for risk assessment. We have also shown that incorporating dosage effects substantially improves the predictive capability of the rules. To make predictions for individual patients, we applied a very simple strategy: for each patient, we estimated his risk of progression using the most specific rule that applies to him. Naturally, building a penalize survival model on top of the survival association rules would be more appropriate for making predictions for individual patients, our goal was to demonstrate that even with this simple strategy, the risk estimates compare well with more flexible survival models, even with state-of-the-art Lasso-penalized survival models. The survival association rules substantially outperformed the popular Framingham score.

References

1. Agrawal R, Srikant R. Fast algorithms for mining association rules. In VLDB Conference, 1994.
2. American Diabetes Association. Executive summary: Standards of medical care in diabetes—2013. In Diabetes Care. American Diabetes Association, 2012.
3. Caraballo PJ, Castro MR, Cha SS, Li PW, Simon GJ. Use of association rule mining to assess diabetes risk in patients with impared fasting glucose. In AMIA Annual Symposium, 2011.
4. Centers for Disease Control and Prevention. National diabetes fact sheet: national estimates and general information on diabetes and prediabetes in the united states, 2011. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention http://www.cdc.gov/diabetes/pubs/factsheet11.htm, 2011.
5. Collins GS, Mallett S, Omar O, Yu LM. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. BMC Medicine, 2011.
6. Diabetes Prevention Program Research Group. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. The New England Journal of Medicine, 346(6), 2002.
7. Friedman JH. Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29:1180, 2001.
8. Kim HS, Shin AM, Kim MK, Kim N. Comorbidity study on type 2 diabetes mellitus using data mining. Korean J Internal Medicine, 27, 2012.
9. Ridgeway G. The state of boosting. Computing Science and Statistics, 31, 1999.
10. Shin AM, Lee IH, Lee GH, Park HJ, Park HS, Yoon KI, Lee JJ, Kim YN. Diagnostic analysis of patients with essential hypertension using association rule mining. Healthc Inform Res, 16(2):77–81, Jun 2010.
11. Simon GJ, Kumar V, and Li PW. A simple statistical model and association rule filtering for classification. In ACM International Conference on Knowledge Discovery and Data Mining (KDD), 2011.
12. Tuomilehto J, Lindstrom J, Eriksson J, Valle T, Hamalainen H , Ilanne-Parikka P, Keinanen-Kiukaanniemi S, Laakso M, Louheranta A, Rastas M, Salminen V, Uusitupa M. Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impared glucose tolerance. The New England Journal of Medicine, 344(18), 2001.
13. Wilson PW, Meigs JB, Sullivan L, Fox CS, Nathan DM, D'Agostino RB. Pediction of incident diabetes mellitus in middle-aged adults–the Framingham offspring study. Archives of Internal Medicine, 167, 2007.