Toward Personalized Care Management of Patients at Risk - The Diabetes Case Study

Hani Neuvirth, Michal Ozery-Flato,
Jonathan Laserson, Michal Rosen-Zvi
Machine Learning and Data Mining group
IBM Research
Mount Carmel, Haifa, 31905, Israel
{hani, ozery, ljon, rosen}@il.ibm.com

Jianying Hu, Martin S. Kohn,
Shahram Ebadollahi
Healthcare Transformation group
IBM Research
IBM T.J. Watson Research, Hawthorne, NY
{jyhu, marty.kohn, ebad}@us.ibm.com

ABSTRACT

Chronic diseases constitute the leading cause of mortality in the western world, have a major impact on the patients' quality of life, and comprise the bulk of healthcare costs. Nowadays, healthcare data management systems integrate large amounts of medical information on patients, including diagnoses, medical procedures, lab test results, and more. Sophisticated analysis methods are needed for utilizing these data to assist in patient management and to enhance treatment quality at reduced costs.

In this study, we take a first step towards better disease management of diabetic patients by applying state-of-the art methods to anticipate the patient's future health condition and to identify patients at high risk. Two relevant outcome measures are explored: the need for emergency care services and the probability of the treatment producing a sub-optimal result, as defined by domain experts. By identifying the high-risk patients our prediction system can be used by healthcare providers to prepare both financially and logistically for the patient needs. To demonstrate a potential downstream application for the identified high-risk patients, we explore the association between the physician treating these patients and the treatment outcome, and propose a system that can assist healthcare providers in optimizing the match between a patient and a physician.

Our work formulates the problem and examines the performance of several learning models on data from several thousands of patients. We further describe a pilot system built on the results of this analysis. We show that the risk for the two considered outcomes can be evaluated from patients' characteristics and that features of the patient-physician match improve the prediction accuracy for the treatment's success. These results suggest that personalized medicine can be valuable for high risk patients and raise interesting questions for future improvements.

Categories and Subject Descriptors

G.3 [**Mathematics of Computing**]: Probability and Statistics – *Survival analysis;* I.2.6 [**Computing Methodologies**]: Artificial Intelligence – *Learning;* J.3 [**Computer Applications**]: Life and Medical Sciences – *health, medical information systems.*

General Terms: Algorithms, Measurement, Performance

KDD'11, August 21–24, 2011, San Diego, California, USA.
Copyright 2011 ACM 978-1-4503-0813-7/11/08...$10.00.

1. INTRODUCTION

Recent advances in adoption of information technology in healthcare organizations have made huge volumes of patient data available. Analysis of this data can uncover patterns for best practices and insight, which can potentially improve care delivery and make medical practices more effective. The measure of goodness of the care delivery process is ultimately obtaining optimal outcomes for the patients while reducing the costs of patient care in the long run.

Chronic diseases are known to consume the greatest majority of healthcare expenditures [4]. Chronic illness extends over many years and is often associated with acute exacerbations and progressive deterioration. One of the goals for healthcare policy is to improve chronic illness treatment and minimize the exacerbations, slow or halt deterioration, and decrease the cost of care. Diabetes is a chronic disease whose incidence is increasing and appearing in progressively younger patients. Moreover, diabetes mellitus is notoriously known for the serious complications associated with long disease duration. Thus, more patients will be dealing with diabetes and its complications for longer periods of time, making it important to improve and evaluate the quality of healthcare provided to diabetics.

An important step towards an improved diabetes treatment is to define a quantitative evaluation measure of the disease status. In this study we focus on two common methods for evaluating diabetes: blood test results and urgent care visits. The HbA1c blood test (also called glycohemoglobin) is a reliable indicator of long term diabetes management. The higher the HbA1c measure, the higher the risk of developing complications such as eye, heart, or kidney disease, nerve damage or stoke. Poorly managed diabetics will need urgent care more often and thus show an increase in the number of emergency department and urgent care visits. The number of these urgent care (UC) events can also serve as a way to measure the status of the disease.

Measurements for the time until an UC event occurs and the HbA1c lab test results can be considered labels or outcome to be predicted in a machine learning or a statistical analysis settings. The estimated probability for these outcomes can help assess the risk for disease aggravation. Determining health risks in the context of particular chronic diseases is very common in the medical literature. The seminal work of Cox in the early seventies [5] provides the most common method for prediction of disease risk and identification of the associated factors. One of the classic examples for the utility of disease risk assessment is cardiovascular disease (CVD). Assessing the risk for CVD has become a major strategy for preventing this disease and its unfavorable consequences. Over the past two decades many prediction algorithms were developed to assess CVD risk (see [6,12] for references). The vast majority of these algorithms use proportional hazards models for their prediction, most commonly the semi-parametric Cox model [5]. A few studies also applied the Cox model to the case of diabetes [3,17]. Commonly, these algorithms use the vast knowledge that has been accumulated on the disease to manually select a limited number of risk factors to be included in the model. Recent studies have proposed using machine learning methods for feature selection and dimensionality reduction combined with Cox regression for better predictions (see [13] and references therein). In this paper, we show that the best performing algorithms are those that combine Cox with various machine learning techniques.

There are very few examples of machine learning applications for personalized healthcare. Among them are prediction of risk of heart attack when consuming particular drugs [7] and prediction of treatment outcome of HIV patients [18]. Our work analyzes longitudinal operational data of patients; the data was gathered during the routine processes of a large health maintenance organization. We explore the use of classic machine learning techniques as well as common statistical methods for survival analysis to achieve the most accurate outcome prediction. Furthermore, we propose ways to use the prediction algorithms to provide better personalized care. This is carried out by employing the prediction models for identifying patients at risk and optimizing the match between those patients and the physicians who treat them to optimize the clinical condition of the patient.

The formulation of the problem includes two potential outcomes and a set of several thousand candidate features. Since the large input data is noisy, we experimented with different methods for reducing the data size, leaving only significant signals that can aid the performance of a classifier. Moreover, it is infeasible for some of the known classifiers, such as Cox, to run on hundreds of features.

There are three main approaches for feature selection. The first includes filters, where each feature is tested against the outcome and an associated score of relevancy or significance of the feature is calculated. The second approach uses wrapper algorithms, which search and select the subset of features that provide the best performance of the particular classifier being used [14]. The third is an embedded approach that incorporates feature selection directly into the learning process of a classifier. A different approach to reduce data dimensionality prior to applying a classifier is to carry out principal components analysis (PCA) of the data and then apply the classifier only on the projected data [23]. We experimented with all these approaches for feature selection and with PCA for dimensionality reduction. As part of our work, we compared the performance of three different classifiers: k-nearest neighbors, logistic regression, and Cox regression. We explored different groups of features and propose methods to increase the accuracy of the prediction by exploiting different feature selection methods and by employing an ensemble method that combines the various algorithms.

This paper describes the different features we derived from the medical data and how they are automatically selected and used to provide accurate predictions of patient status. The prediction includes both visits to urgent care departments in the following year and treatment outcome based on HbA1c results in follow-up tests. The ability to predict the future status of a chronic patient a year in advance is especially valuable for health organizations that need to manage and estimate expected expenses associated with different patients. The main contribution of this paper is the integration of the above methods into an innovative application that provides accurate predictions on the future condition of a chronic patient. This prediction has the potential to help health organizations optimize and tailor the care they provide to the individual needs of the patients.

2. METHODS

2.1 Data Organization and Problem Formulation

Our data comprises four main sources: claims, pharmacy, labs, and patient profiles, covering around 200,000 patients over a three year period. All sources have been de-identified and a unique patient id is available to link them. The key data fields in each source are:

- Claims: Patient ID, Date of Service, Diagnosis Code (ICD9), Procedure Code (CPT), ID of Physician Seen, Specialty of Physician Seen, Facility Type, Cost.
- Pharmacy: Patient ID, Date Filled, Drug Code (NDC), Dosage, Days of Supply.
- Labs: Patient ID, Test Name, Date of Test, Test Result, Unit.
- Patient Profile: Patient ID, Age, Gender, Assigned primary care provider (PCP)

Various data mapping steps were applied to make the data more amenable for extracting clinically meaningful features. For example, ICD-9 codes presented a challenge because diagnoses that are logically related in terms of risk-adjustment or clinical syndromes are often not indicated as such simply by the ICD-9 code. Consequently, several different diagnostic categorizations were employed including: HCUP-US Clinical Classifications Software (CCS) for Diagnosis (both hierarchical and single-level forms) and Hierarchical Diagnostic Categories (HDC), an IBM customized, extended variant of the Hierarchical Condition Categories (HCC) used in Medicare Risk Adjustment. Similarly, the CPT procedure codes were mapped to CCS for procedures. The Relative Value Unit (RVU) for each procedure, as defined in the Resource-Based Relative Value Scale (RBRVS) [10] is retrieved and used as an indicator of the effort level. Finally, the National Drug Code (NDC) Directory, obtained from FDA, was utilized to cross-index each NDC to its corresponding generic drug components.

2.1.1 Samples definition

For our investigation, a patient is considered to have diabetes if she had at least one hemoglobin A1c (HbA1c) test result that is higher than 6.5 [15]. Although a diagnosis of diabetes can be readily determined from the diagnosis codes, we found that such codes are not always reliable. One reason for the lack of reliability is that sometimes a patient is given a diagnosis code corresponding to diabetes to justify the prescription of an HbA1c test, even though the patient will ultimately not be diabetic.

For each patient determined to have diabetes, the day after the earliest HbA1c test result is selected as the *reference date* for the patient (Fig. 1). This reference date is used to separate the pre-reference condition evaluation period (at least 90 days), from the post-reference treatment period. Samples are further filtered to meet a requirement on the minimum number of patients treated by a physician, and on the duration of available data in the two periods.

Many patients in our data were being monitored for less than the defined follow-up period of one year. For example, consider the case in which the predicted outcome is the occurrence of an urgent care event. A patient for whom the event already occurred can be regularly analyzed. For the other “short-term” patients we have only partial information: we know that the event did not occur till a certain time. We refer to such partial data as *censored data*.

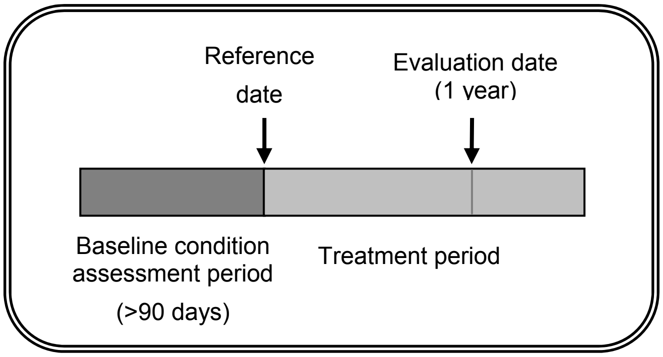


Figure 1: Formulation of a patient's longitudinal data

2.1.2 Labeling based on domain expert input

Two outcome measures are explored in this study: the occurrence of urgent care, as a surrogate measure of the degree of success in avoiding severe complications; and the control of HbA1c level, which is the outcome measure defined in most diabetes care quality metrics including the Comprehensive Diabetes Care criteria defined by National Committee for Quality Assurance (NCQA) [2].

For the first outcome measure, we checked the “Facility Type” field in the claims records. An encounter is considered emergency care if its facility type is either “Urgent Care” or “Emergency Room Hospital”. In our data set, these visits account for about 4% of the total encounters. An outcome is labeled as "desirable" if the patient has no recorded emergency care encounter as defined above during the one year treatment evaluation period; otherwise it is labeled as "undesirable".

For the second outcome measure, we used widely adopted clinical ranges defined as follows [15]: HbA1c ≤6.4: normal; 6.5 ≤ HbA1c < 7: well controlled; 7≤HbA1c<9: moderately controlled and 9 ≤HbA1c: poorly controlled. Each patient’s treatment outcome is labeled by comparing the last HbA1c test result in the condition evaluation period with the one obtained after one year of treatment (allowing up to 2 months deviation). The outcome is labeled "desirable" if the HbA1c level moved into a lower range, or remained in the well controlled range, otherwise it is labeled as "undesirable".

2.1.3 Feature construction

The patient features derived from the database were used to represent patients' baseline condition. The first two features are the age and sex of the patients. Most other features are counts of records grouped by different criteria. These include the total and distinct number of CPT medical procedure codes grouped by their associated CCS code; the number of ICD-9 diagnoses codes grouped by their CCS diagnostic categorization, weighted by their HDC risk measure, or grouped and weighted by several risk categories defined by Young et al.[24]; the number of drug prescriptions and dosage for each ingredient and for each NDC drug code; the number of times each lab test was performed, and the number of visits to each facility type. Another feature weighted each CCS medical procedure category by its total effort invested for each patient as evaluated by the RVU. An advantage of all these features for medical data is that by definition, zero is a valid value, thus serving as a natural default value for cases in which some features were irrelevant for some of the patients based on medical considerations (e.g., lab tests that were not performed, as the medical condition did not require them). This also implies that many of the features are highly sparse.

In addition to the above, lab tests were also represented by the fraction of lab tests that are out of the optimal range for that patient and by the mean value of the lab test. For the latter, missing data had to be imputed. In the training and test sets, we replaced these data by the mean lab test values calculated on the training set alone.

Three types of features were derived. Patient features are described above. These features were calculated for the period prior to the reference date and were used to characterize a patient's baseline condition (Fig. 1). This set of features included 3236 different features that are used in the two analyses presented in the paper – for high-risk patient identification and for the patient-physician outcome modeling.

The features we used to characterize physicians were based on the population of their patients. These features aim to characterize a physician's treatment, and thus were derived from the period following the reference date. They were calculated by taking the mean of patient features separately on the two sets of patients that achieved desirable and undesirable outcomes. The corresponding patient for which the prediction was being performed was excluded from the mean. If one of the desirable or undesirable population set of a physician was empty, the mean over the corresponding population for the whole set of physicians was used.

The third group of features we used is patient-to-physician (P2P) features, which characterizes the match of the patient and the physician by quantifying the similarity of the patient to the characteristic population of the physician. These were calculated over the period preceding the reference date by taking the distance between patient's features and each of the two classes' population means of the physician. These means were calculated similar to the set of the physician's features above, on the baseline condition evaluation period.

To formalize the above, let $f_{pat,pre}(i), f_{pat,post}(i)$ denote a specific patient-feature of patient *i*, for the periods before and after the reference date, respectively. Let $G^{+}(j), G^{+}(j)$ denote the groups of patients treated by a physician *j* after the reference date, and who in practice achieved a desirable and undesirable outcome respectively. Then, we define the positive physician features as:

$$f^{+}_{phys,post}(i,j) = \frac{1}{|G^{+}(j)-\{i\}|}\sum_{k\in G^{+}(j)\}\setminus\{i\}} f_{pat,post}(k)$$

the distance between the patient and the positive set of physician patients as:

$$d^{+}_{P2P}(i,j) = | f_{pat,pre}(i,j) - f^{+}_{phys,pre}(i,j) |$$

the positive patient-to-physician features as:

$$f^{+}_{P2P}(i,j) = \exp\left(-\frac{d^{+}_{P2P}(i,j)}{Var_{i',j'}(d^{+}_{P2P}(i',j'))} \right)$$ and

correspondingly for the negative feature types derived from G_j^{-} .

2.2 Performance Evaluation Measures

We used two standard outcome measures to evaluate the results. The area under the receiver operating characteristic curve (AUC), and the concordance index (c-index). The c-index is a common performance measure in survival analysis; it evaluates the correspondence between the prediction and the survival times [9]. It is derived from the Willcoxon-Mann-Whitney two-sample rank test and can be regarded as an extension of the AUC to censored data. The interpretation of both measures is similar, with values around 0.5 denoting no significant predictive power and 1 indicating perfect prediction. We evaluated the performance as the mean and standard deviation on 10 different 5 fold cross-validation partitions of the dataset.

2.3 Learning Models

We evaluated and compared different analysis models following two approaches: binary classification and survival analysis. We explored two common binary classification algorithms: logistic regression (LR), and k-nearest neighbor. The survival analysis was performed with the Cox proportional hazards model. Finally, we also evaluated ensembles of the methods using a combination rule taking the mean of all predictors and the two best performing ones.

Logistic regression is a popular binary classification algorithm aiming to model the posterior probabilities of the two classes. It belongs to the classes of generalized linear models with a logit link function and a binomial distribution. Thus, the predicted outcome is the probability of a sample belonging to each class.

K-nearest neighbor (kNN) is another algorithm often used in machine learning applications. In a classification setting each test sample is classified based on its k nearest neighbors from the training set. Several variations of this algorithm exist. We used a weighted approach where the classification of a sample was a weighted mean of the k nearest neighbors with exponentially decaying weights. Specifically,

$$\hat{y}(i) = \frac{\sum_{j=1}^{k} \exp(-dist(i,j)) * y(j)}{\sum_{j=1}^{k} \exp(-dist(i,j))}$$

with *dist* denoting the Euclidean distance.

The Cox model is designed for longitudinal data in which the outcome is a pair (δ, t), where δ is an indicator for an event and t is the time in which δ was observed. In our study we formulated care services data for patients in the following way: δ =1 indicates that the patient required a care service in the following year, and t is the first time such service was given; δ=0 indicates that the patient required no service while being under study, and t is the overall time this patient has been under study.

Consider a random variable T ∊ R^+ representing the time till the event. The survivor function is defined by S(t) = P(T≥t). The Cox model is based on an assumption that given the features vector, **x**, the log of the survivor function can be expressed by the following equation [16]:

$$\log S(t \mid x) = \exp(\beta^T x) \log S_0(t)$$

where **β** is a vector of coefficients in the size of features vector **x**. This model is considered semi-parametric as the functional form of the baseline survivor function, $S_0(t)$, is not given, but determined from the data.

We trained all learning models using 5 fold cross-validation. The entire learning flow was performed over the training set, starting from data imputation, through the feature selection and the final optimization of the learner. We performed the corresponding adjustments on the test set (data imputation, feature selection and normalization) based on the training set values.

2.3.1 Feature selection

We performed two baseline feature selection procedures for all models to remove uninformative and redundant features. We removed sparse features in which 99% of the patients have identical values or the standard deviation was below 1e-6. We calculated the correlation coefficient (CC) for each pair of features and discarded one of the features in each highly correlated pair (CC > 0.9).

Next, we employed one of four feature selection and dimensionality reduction strategies:

- *Filter:* This feature selection approach is a fast pre-processing step, performed independently of the learning model. We adapted the feature selection for each model, thus binary classification methods were preceded by filters based on the p-values coming from one of two standard statistical tests: the chi square test for discreet features and the Kolmogorov-Smirnov test for continuous features, detecting features with distinguished distributions between the two classes. For the Cox regression we tested the t-test p-values for each coefficient to be different from zero in a model consisting of this single feature. We then filtered out features whose corresponding p-values were insignificant.
- *Wrapper*: This mode of feature selection is a meta-algorithm that evaluates the performance of the model under different selection of features. In this work we enhanced the Cox model by using forward stepwise feature selection with Akaike information criterion (AIC) for model evaluation.
- *Embedded optimization:* Under this scheme the feature selection is an integral part of the model. We used lasso [20], a popular method for regression that estimates the vector of coefficients, β, under a constraint that bounds $||\beta||_1$ by a constant. This method results in the shrinkage of coefficients, setting some coefficients to exactly zero, thus yielding a sparse solution. This method has been extended for generalized linear models (including logistic regression) [20] and the Cox model [21]. In our analysis we used the implementation available in the glmnet R package [8].
- *PCA:* This is a classic dimensionality reduction scheme. Features are projected on a set of orthogonal axes selected such that the first axis is aligned with the maximal variance of the data, the second axis is orthogonal to it, and is aligned

with the second largest variance axis, and so forth. Thus, selecting a subset of these axes can preserve most of the variance in the data while reducing the dimension of the data. In this work, we reduced the features to cover 90% of the variance in the data.

3. RESULTS

3.1 Data Statistics

Three datasets have been derived for each outcome. The first {Physician:All, Patients:All} is the full dataset of diabetic patients, used to construct a prediction model for patients at risk for an undesirable outcome. The other two datasets {Physician: >5 patients} are used for the secondary analysis of patient-physician match, and thus, in order to have sufficient statistics per physician, a requirement on the minimal number of patient attending each physician was added. The dataset of all the patients with the subset of the physicians seeing at least 5 patients was used to re-evaluate the patients at risk. Then, the top high-risk patients were selected to train and evaluate the patient-physician outcome model. Two contradicting considerations guided us in choosing the size of the top high-risk patients. First was to have enough information to train and evaluate the patient-physician outcome model, and second was to obtain a well characterized high-risk population. For the HbA1c outcome we selected the top 100 patients. Since in the UC model about half of the data is censored, in order to have enough samples, the top 200 patients were selected. For validation purpose we preserved the same cross-validation partition in the primary analysis identifying patients at risk and in the secondary analysis of patient-physician outcome prediction on the selected patients' subset. A more rigorous analysis would have used two embedded cross-validation partitions.

Table 1 presents the sample counts of the above datasets, providing detailed counts for the two classes of each outcome. In the UC context, the desirable outcome is the patient *not* attending emergency care services. In the HbA1c context, the desirable outcome is a reduced level of HbA1c. (See methods for details.)

Table 1. Sample counts in each of the datasets used

Outcome	**Urgent care**			**HbA1c**		
Physician	**All**	**> 5 patients**		**All**	**>5 patients**	
Patient	**All**	**All**	**At risk**	**All**	**All**	**At risk**
Desirable	2400	2374	40	1586	1412	43
Undesirable	541	531	78	666	608	57
Censored	1604	1550	82	--	--	--
Total	4545	4455	200	2152	2020	100

For some patients, censoring of UC information was before the date of the HbA1c lab test. Therefore, these individuals appear in the HbA1c dataset, but not in the UC dataset. The overlap between the datasets of the two outcomes consists of 2003 individuals, with 1462 individuals overlapping when censored data is excluded. No significant correlation was found between the two outcomes, as calculated on this overlapping subset. Thus these measures reflect complementary aspects of the disease management.

3.2 Identifying Patients-at-Risk

We compared the prediction accuracy achieved by various learning models on the two outcomes used to define patients at risk. The UC outcome is the more acute one, targeted at patients who reach emergency conditions. The common formulation of this type of data is under the framework of survival analysis. The most abundant model in the literature for this type of analysis is the Cox proportional hazards model [5]. This model is capable of handling events' time information, as well as censored data.

Table 2. Prediction performance of the high-risk patients

Outcome	**Method**	**No. Features**	**C-Index**	**AUC**
UC	Cox + LASSO	46 (±7)	0.52 (±0.07)	0.57 (±0.01)
UC	kNN(10) + filter(0.05)	41 (±1)	0.574 (±0.009)	0.63 (±0.01)
(c) UC	kNN(100) + filter(0.05)	41 (±1)	0.634 (±0.006)	0.669 (±0.008)
UC	Cox + filter(0.05)	168 (±3)	0.645 (±0.008)	0.671 (±0.009)
UC	Cox + filter(0.05) + PCA	108 (±2)	0.652 (±0.007)	0.679 (±0.008)
UC	Cox + filter(0.001) + stepwiseAIC	26 (±1)	0.652 (±0.004)	0.679 (±0.005)
UC	Cox + filter(0.001) + PCA	35 (±1)	0.653 (±0.003)	0.680 (±0.005)
(a) UC	Cox + filter(0.001)	51 (±2)	0.659 (±0.003)	0.685 (±0.004)
(b) UC	LR + filter(0.05)	41 (±1)	0.666 (±0.003)	0.713 (±0.003)
UC	Ensemble (a,b,c)	179 (±3)	0.670 (±0.003)	0.713 (±0.004)
UC	Ensemble (a,b)	71 (±2)	0.672 (±0.002)	0.714 (±0.003)
HbA1c	LR + filter(0.05)	62 (±3)	--	0.725 (±0.004)

An alternative approach we examined here for the same outcome was to formulate the data as a binary classification problem. The class labels were based on whether the patient attended UC services in the first year after the reference date. Thus, the data available for the classification approach excluded all censored samples and included only those samples having full information about the first year. This excluded 1604 samples, which represent 35% of the dataset. Also, the specific UC events time information was ignored.

Despite the limited and less informative training data available for the binary classification formulation, the logistic regression model outperforms the Cox proportional hazards model (Table 2). This is consistently reflected by both the AUC, which is calculated only on the non-censored data, and the c-index, which includes the censored data in the evaluation. This occurred despite the fact that the censored data constituted more than a third of the dataset.

One difference stands out when comparing the two classic models. Using the filter feature selection test appropriate for each algorithm with a bound of 0.05 on the p-value, the LR filter excludes significantly more features than the Cox filter. While the LR model is trained on 41 features on average, the Cox proportional hazards model is trained on a set of features about four times larger, and thus the risk of overfitting is significantly higher. Adapting the bound on the p-value to 0.001, which resulted with a similar number of features as for the LR, significantly improved the results.

To make sure the results are not biased by the different training set used, and to discard the option that the binary classification data are better in some sense, we also trained the Cox model with the 0.001 filter on the limited dataset available for the classification algorithm. This model achieved a c-index of 0.642±0.006 and an AUC of 0.526±0.006, suggesting that the utilization of the data by the LR flow model is superior.

Several enhancements of the Cox proportional hazards model have been developed in recent years. These are mainly based on sophisticated feature selection techniques. As an alternative to the naïve filter, we employed an embedded feature selection approach with the lasso optimization for Cox, and a wrapper approach running sequential feature selection to optimize the Akaike information criterion (AIC). As the performance of previous models improved when the number of parameters was 40-50, the parameter of lasso was configured to choose 50 features[1]. The stepwise feature selection was preceded with Cox-filtering with p-value bound of 0.001 in order to reduce the number of features, since running this method is infeasible when the number of features is large. Both approaches fail to improve on the simple filter approach with a p-value bound of 0.001.

An alternative strategy for reducing the number of features while preserving the information is to use dimensionality reduction techniques. PCA is one such simple, state-of-the-art technique. We employed PCA on the filtered features keeping the principal components covering 90% of the variance in the data. Using this approach the performance of the model with 0.05 feature selection bound was improved, however, no improvement on top of the strict feature selection model was observed.

The success of the logistic regression model encouraged us to explore the performance of another binary classification algorithm – k nearest neighbor. We present here the results of a variant of the algorithm using a weighted average of the neighbors' classification with weights exponentially decaying with the distance from the query point. The performance of this approach with 10 and 100 neighbors was unsatisfactory. Testing the unweighted variant completely failed learning (data not shown).

Different learning models that are trained on the same dataset can capture different properties of the data, and thus their error regions might differ. In those cases, an ensemble of the models may perform better than each of the individual models. We used a simple combination rule taking the mean of all three approaches, and another ensemble of only the two winning models (the Cox proportional hazards, and the logistic regression). The resulting prediction accuracy was slightly improved.

The HbA1c outcome is a less extreme outcome measure based on domain-experts' evaluation of the treatment quality; it compares the blood sugar levels measured by the hemoglobin A1c lab test at the reference date and 1 year ± 2 months later. The prediction accuracy for this outcome is of the same order of magnitude as the UC events. It would be interesting to compare the performance of various applications that can be utilized for these different sets of high-risk patients derived by these two different measures.

Table 3. A representative set of features that were consistently selected in 10 repetitions of 5-fold cross-validation (50 times) by the HbA1C and UC models

feature class	Feature description	HbA1C	UC
lab tests	HbA1c	50	50
	LDL	50	7
	HDL	50	14
	Cholesterol, Total	50	38
	Triglycerides	50	45
	Microalbumin/Creatinine Ratio test count	50	0
	Triglycerides test count	50	5
	% of out-of-range lab tests	50	13
drugs	Number of Insulins prescriptions	50	0
	Number of Antiypergycemic prescriptions	50	2
	Number of prescriptions for the ingredient Glyburide	50	0
diagnoses	Total HDC weighted ICD-9 codes associated with Diabetes mellitus with complications	50	0
	Number of ICD-9 codes associated with: Other diagnostic radiology and related techniques	50	29
	Number of distinct ICD-9 codes associated with: Other upper respiratory infections	0	50
	Number of distinct ICD-9 codes associated with: Other nutritional; endocrine; and metabolic disorders	1	50
medical procedures	Number of procedural codes: Microscopic examination	49	50
facility	Emergency Room Hospital	0	50
	Urgent Care	0	50
	Hospital – Outpatient	50	47

3.2.1 Selected features

Based on the results of the different high-risk identification models, we decided to use the logistic regression model to predict

[1] We used the 'glmnet' R package. which allows specifying a bound on the number of features. This may yield a selection of less features than indicated by the given bound.

UC and HbA1c outcomes in all subsequent analyses. To learn about the features being selected, we checked the consistency of the selection over the 10 repetitions of 5-fold cross-validation. Thus, every feature can potentially be selected at most 50 times by each model. We examined the features repeatedly selected by all the models. These included 16 features for the UC model and 31 features for the HbA1c model, with 6 features overlapping. Table 3 shows a representative set of 19 of these features. Many of the selected features are associated with known diabetes risk factors or consequences of the disease and its identification, supporting the clinical relevance of our models. The baseline hemoglobin A1c feature, which is the long term tracking parameter for diabetes, was selected by the two models. This was expected for the HbA1c model, which predicts the improvement in this parameter. However, it was also found significant with the UC outcome, corroborating the correlation between the severity of the diabetes and the UC events. Other features selected by the two models include the total cholesterol and triglycerides - lab tests measuring blood fats level. In particular, abnormalities in triglycerides level are known to be associated with diabetes.

The difference between the two outcomes becomes apparent when considering the differences at the feature class level. The features related to lab tests, diagnoses, and medical procedures are in agreement. The main difference lies in the drugs features, which are more significant for the HbA1c outcome, and the facility features, which have a bigger weight in the UC outcome. Indeed, it is expected that drugs would have a more significant effect on the relevant lab test, while UC events will be indicative of future events. Hence, these two feature classes are better tailored for each of the specific outcomes.

Examining the differences at the features level, those that were selected by the HbA1c model but not by the UC model include: Microalbumin/Creatinine Ratio test count - a lab test indicating diabetic kidney disease in case of abnormal values; Total HDC weighted ICD-9 codes associated with the CCS category of "Diabetes mellitus with complications" - increase in this parameter implies progression of the disease; Number of prescriptions for insulins, antihyperglycemic drugs, and glyburide – all of them are diabetic medications implying a longer history of diabetes. Features that were selected by the UC model but not by the HbA1c model include UC and emergency room visits during the baseline evaluation period, before the reference date, and the number of ICD-9 codes related to diagnoses of upper respiratory infections. These features imply poor disease management of the patient.

3.3 Patient-Physician Outcome Prediction

The first part our work shows how learning algorithms can be used to identify patients at high risk of failing their treatment. We claim that the treatment quality of these patients can be significantly enhanced by taking a personalized healthcare approach. This section presents one downstream analysis demonstrating the value of our approach. The analysis aims to predict the best physician assignment for each of the high-risk patients, based on the specific characteristics of the patient, the physician, and their match.

Similar to the high-risk patient identification analysis, patients were represented using the features derived directly from the database. Physicians were represented using the population of all their patients, excluding the index patient. These features were derived separately for their patients that reached a desirable outcome and those whose treatment failed. The patient-to-physician (P2P) match was represented by the similarity of the index patient to each of the two characteristic populations of the physician.

Assuming a hypothetical physician assignment reference date, patient and P2P features were derived from the period preceding this date, while the features characterizing the physician's treatment were derived from the period following the reference date. The labels for each patient-physician match were evaluated a year after the assignment date (see methods).

Table 4. Prediction performance of patient-physician match with the urgent care outcome

Feature Sets Used			C-Index		AUC	
Pat.	Phys.	P2P	mean	std	mean	std
0	1	0	0.51	0.06	0.52	0.11
1	1	0	0.54	0.05	0.48	0.10
1	0	0	0.56	0.04	0.48	0.05
0	0	1	0.60	0.02	0.53	0.05
0	1	1	0.60	0.02	0.53	0.05
1	1	1	0.61	0.03	0.53	0.05
1	0	1	0.61	0.02	0.54	0.05

Table 5. Prediction performance of patient-physician match with the HbA1c outcome

Feature Sets Used			AUC	
Pat.	Phys.	P2P	mean	std
1	0	0	0.45	0.02
0	1	0	0.49	0.06
1	1	0	0.49	0.06
1	0	1	0.56	0.03
0	0	1	0.56	0.03
1	1	1	0.59	0.03
0	1	1	0.59	0.03

Tables 4 and 5 present the prediction accuracy for the two outcomes used, with all possible combinations of feature classes. Having 1 in the feature set's column indicates it was used in the model, and 0 means this feature set was not used. Naturally, the patient model does not provide additional information on the treatment quality. However, features characterizing the physician and patient-physician match can further enhance the prediction up to a c-index of 0.61 ± 0.02 for the UC outcome and an AUC of 0.59 ± 0.03 for the HbA1c outcome. This implies that for this subset of patients, the personalized physician assignment plays a role in the treatment's success. In a similar analysis of the full patient population, the winning model was the one based solely on the patient features; adding physician characteristics did not affect the model's performance (data not shown).

Figure 2 presents the predicted outcome for the top 100 high-risk patients based on the HbA1c outcome. The top part of the panel shows the patients who in practice achieved a desirable outcome, while the bottom part represents patients whose treatment failed. One can see the variance of the expected outcomes of different physicians on different patients. Physicians on the right are expected to succeed for all patients, while physicians in the left columns fail more often. In addition to providing a data-driven physician assessment tool, this calls for further research on exploring the differences in treatment strategies among different physicians. A similar figure can be produced to present the predictions for the UC outcome.

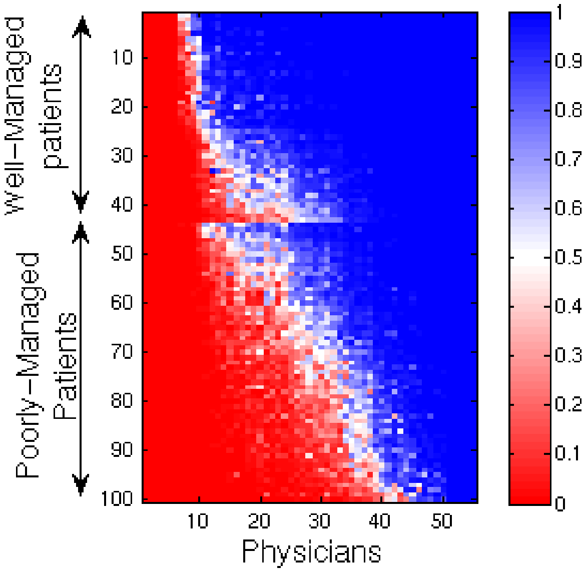
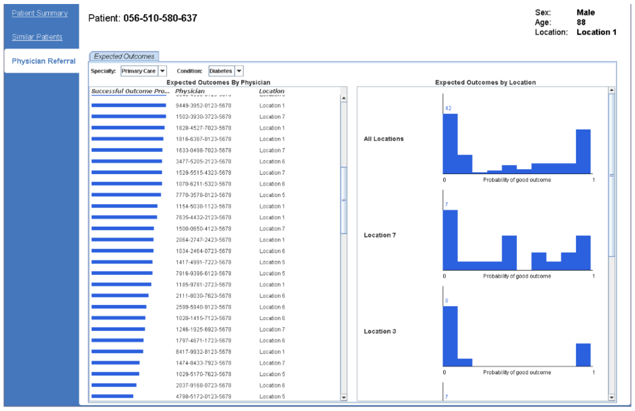


Figure 2. Predictions of treatment success probability for the high-risk patients with all possible physicians based on the HbA1c outcome. Desirable predicted outcomes appear in blue, and undesirable predicted outcome appear in red.

4. SYSTEM DEMO

The methodologies described in this paper were incorporated into a pilot system as part of a suite of analytics tools designed to support personalized decision making for both care delivery and practice management, particularly in a collaborative care or patient centered medical home (PCMH) setting. PCMH is an enhanced care model that provides comprehensive, coordinated, and timely care and payment reform, emphasizing the central role of primary care physicians [11].

Figure 3 shows a screen shot from this pilot system. The screen captures the output of the patient-physician outcome model described above for a specific index patient, and summarizes it using a sorted histogram visualization. Each potential physician is listed in a table along with a graphical bar whose length represents the predicted probability of achieving a desirable outcome for the corresponding physician. The expected outcome information is further grouped by practice location on the right side of the panel. Unlike many existing physician quality measures that evaluate physicians at aggregate levels [2], our models provide nuanced and personalized information regarding which physicians are good at managing each specific patient. In other words, the information shown on the screen is unique to each individual patient. Such a tool can be used by the medical director of a care organization to better manage high-risk patients. For example, it can be used to determine the physician or practice to which a poorly managed patient should be directed. It can also be used to guide investigations into what certain physicians do that makes them particularly successful in handling a specific type of patient. Then proper education or payment incentive programs can be put in place to improve the outcome for high-risk patients.



Patient: **056-510-580-637**

Sex: **Male**
Age: **88**
Location: **Location 1**

Specialty: Primary Care
Condition: Diabetes

Expected Outcomes By Physician

Successful Outcome Pro...	Physician	Location
	9449-3952-0123-5678	Location 1
	1502-3930-3723-5678	Location 7
	1828-4527-7023-5678	Location 1
	1816-6387-8123-5678	Location 1
	1633-0498-7023-5678	Location 7
	3477-5205-2123-5678	Location 6
	1529-5515-4323-5678	Location 7
	1070-6211-5323-5678	Location 6
	7770-3578-0123-5678	Location 5
	1154-5038-1123-5678	Location 1
	7635-4432-2123-5678	Location 1
	1500-0650-4123-5678	Location 7
	2064-2747-2423-5678	Location 1
	1034-2464-0723-5678	Location 6
	1417-4991-7223-5678	Location 5
	7916-9396-6123-5678	Location 5
	1185-9781-2723-5678	Location 1
	2111-8030-7623-5678	Location 6
	2599-5940-9123-5678	Location 6
	1020-1415-7123-5678	Location 6
	1246-1925-6923-5678	Location 7
	1797-4671-1723-5678	Location 6
	8417-9932-8123-5678	Location 1
	1474-8433-7923-5678	Location 7
	1029-5170-7623-5678	Location 5
	2037-9168-0723-5678	Location 6
	4798-5172-0123-5678	Location 5

Expected Outcomes by Location

All Locations
42
0
Probability of good outcome
1
Location 7
7
0
Probability of good outcome
1
Location 3
8
0
Probability of good outcome
1
7

Figure 3. Screenshot of a decision support tool incorporating patient physician outcome prediction

5. DISCUSSION

This paper presents a prototype for a data-driven risk assessment system for patients with chronic diseases and its novel application for identifying physicians who can deliver optimal care to those patients. The prototype was designed and implemented for the use case of diabetes. It was evaluated on the routine operational data from approximately 4500 diabetic patients; this data was obtained from a large health maintenance organization. The prototype was incorporated into a pilot for healthcare management system that supports personalized decision making. Snapshots of the system demo, similar to the one that will be deployed for the client, are provided.

The paper describes the complete analysis process and addresses the numerous challenges encountered on the way, from data organization and feature construction to model selection and evaluation. The choices made at each of these steps may potentially have a large effect on subsequent steps and the overall system performance. One of the initial challenges faced when designing a research on medical data is defining the outcome. In this study we explored two alternative measures for a patient's outcome. These measures are fundamental in the evaluation of treatment quality for diabetes patients, and fall under two types of analytic frameworks: classification and survival analysis. Additional outcome measures should be considered in the future. The prototype system allows the end-user to interactively select the desired outcome. We tested and compared a variety of learning approaches, including the Cox proportional hazards model, a statistical method commonly used for analyzing longitudinal data, and several state-of-the-art machine learning methodologies.

Analyzing clinical records data is a highly challenging task. Part of the challenge arises from the sizable data involved, its considerable number of patients, and the vast number of features that can be derived. A second aspect is that some of the data is censored, thus providing only partial information about these patients. Moreover, missing and noisy data further increases the difficulty. Our work explored how modern feature selection approaches can be used to enhance the classic Cox proportional hazards algorithm. Through this process, we increased the analysis performance by more than 4% from 0.645±0.008 to 0.672±0.002; this is a significant amount considering the size of our data.

On our data, binary classification methods and specifically logistic regression achieved better results. This is despite the fact that censored data and time information were not utilized by the logistic regression. Since censored data may contain valuable information, there is a real need to develop new techniques that extend machine learning algorithms to handle censored data, such as those made for SVM [19,22].

The results of the personalized patient-physician match application demonstrate the advantage in focusing on high-risk patients. While chronic disease management usually follows well defined guidelines [1], these guidelines are designed to fit the majority of the population. High-risk patients are the ones for whom these guidelines may not be well-suited, thereby increasing the chance that their treatment will fail. Personalized healthcare systems that help identify these exceptional patients can eventually enhance our knowledge of the disease, particularly in the extreme edges of the population, and lead to higher quality medical care. The use of this application has the potential to further improve healthcare practice management. For example, novel financial incentive mechanisms could be adopted to improve the utility of the resources in the care delivery network to drive optimal outcomes for the patients. In conclusion, investigating patient-physician properties can reveal the patterns of best practice adopted by the high performing physicians. Such best practice patterns can then be used to educate the physician population to better treat patients with similar characteristics.

6. ACKNOWLEDGMENTS

We thank Dr. Lavi Shpigelman for fruitful discussions. Michal OF and Michal RZ were partially supported by Lithuanian Research Centre LTD.

7. REFERENCES

[1] Executive Summary: Standards of Medical Care in Diabetes—2011. Diabetes Care December 30, 2010 vol. 34 no. Supplement 1 S4-S10

[2] HEDIS 2011 Technical Specifications for Physician Measurement (Print edition), NCQA.

[3] Andersen, P.K. et al. 1985. A Cox Regression Model for the Relative Mortality and Its Application to Diabetes Mellitus Survival Data. *Biometrics*. 41, 4 (Dec. 1985), 921-932.

[4] Bodenheimer, T. et al. 2009. Confronting The Growing Burden Of Chronic Disease: Can The U.S. Health Care Workforce Do The Job? *Health Affairs*. 28, 1 (Jan. 2009), 64 -74.

[5] Cox, D.R. 1972. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*. 34, 2 (Jan. 1972), 187-220.

[6] D'Agostino Sr, R.B. et al. 2008. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation*. 117, 6 (2008), 743.

[7] Davis, J. et al. 2008. Machine Learning for Personalized Medicine: Will This Drug Give Me a Heart Attack? *the Proceedings of International Conference on Machine Learning (ICML)* (2008).

[8] Friedman, J. et al. 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*. 33, 1 (2010), 1.

[9] Harrell, F.E. 2001. *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. Springer Verlag.

[10] Hsiao, W.C. et al. 1988. Estimating physicians' work for a resource-based relative-value scale. *New England Journal of Medicine*. 319, 13 (1988), 835–841.

[11] J. Adams, P. Grundy, M. S. Kohn, and E. L. Mounib, Patient-Centered Medical Home - What, why and how? [Online] Available: ftp://public.dhe.ibm.com/common/ssi/ecm/en/gbe03207usen/GBE03207USEN.PDF

[12] Jackson, R. et al. 2005. Treatment with drugs to lower blood pressure and blood cholesterol based on an individual's absolute cardiovascular risk. *The Lancet*. 365, 9457 (2005), 434–441.

[13] Khosla, A. et al. 2010. An integrated machine learning approach to stroke prediction. *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (New York, NY, USA, 2010), 183–192.

[14] Kohavi, R. and John, G.H. 1997. Wrappers for feature subset selection. *Artificial intelligence*. 97, 1-2 (1997), 273–324.

[15] LabCorp: Test Menu. *https://www.labcorp.com/wps/portal/provider/testmenu*.

[16] Lawless, J.F. 2002. *Statistical Models and Methods for Lifetime Data*. Wiley-Interscience. Chapter 7.

[17] Perkins, B.A. et al. 2003. Regression of microalbuminuria in type 1 diabetes. *The New England Journal of Medicine*. 348, 23 (Jun. 2003), 2285-2293.

[18] Rosen-Zvi, M. et al. 2008. Selecting anti-HIV therapies based on a variety of genomic and clinical factors. *Bioinformatics*. 24, 13 (2008), i399.

[19] Shivaswamy, P.K. et al. 2008. A support vector approach to censored targets. *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on* (2008), 655–660.

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

[21] Tibshirani, R. 1997. The lasso method for variable selection in the Cox model. *Statistics in Medicine*. 16, 4 (1997), 385–395.

[22] Van Belle, V. et al. 2011. Improved performance on high-dimensional survival data by application of Survival-SVM. *Bioinformatics*. 27, 1 (2011), 87.

[23] Wood, F. et al. 1987. Principal component analysis. *Chemometr. Intel. Lab. Syst*. 2, (1987), 37–52.

[24] Young, B.A. et al. 2008. Diabetes complications severity index and risk of mortality, hospitalization, and healthcare utilization. *AMERICAN JOURNAL OF MANAGED CARE*. 14, 1 (2008), 15.