Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
0(0) 1–16

DOI: 10.1177/0962280218810911
journals.sagepub.com/home/smm
SAGE

Predicting diabetes-related hospitalizations based on electronic health records

Theodora S Brisimi, Tingting Xu, Taiyao Wang, Wuyang Dai and Ioannis Ch Paschalidis

Abstract

Objective: To derive a predictive model to identify patients likely to be hospitalized during the following year due to complications attributed to Type II diabetes.

Methods: A variety of supervised machine learning classification methods were tested and a new method that discovers hidden patient clusters in the positive class (hospitalized) was developed while, at the same time, sparse linear support vector machine classifiers were derived to separate positive samples from the negative ones (non-hospitalized). The convergence of the new method was established and theoretical guarantees were proved on how the classifiers it produces generalize to a test set not seen during training.

Results: The methods were tested on a large set of patients from the Boston Medical Center – the largest safety net hospital in New England. It is found that our new joint clustering/classification method achieves an accuracy of 89% (measured in terms of area under the ROC Curve) and yields informative clusters which can help interpret the classification results, thus increasing the trust of physicians to the algorithmic output and providing some guidance towards preventive measures. While it is possible to increase accuracy to 92% with other methods, this comes with increased computational cost and lack of interpretability. The analysis shows that even a modest probability of preventive actions being effective (more than 19%) suffices to generate significant hospital care savings.

Conclusions: Predictive models are proposed that can help avert hospitalizations, improve health outcomes and drastically reduce hospital expenditures. The scope for savings is significant as it has been estimated that in the USA alone, about $5.8 billion are spent each year on diabetes-related hospitalizations that could be prevented.

Keywords

Diabetes mellitus, hospitalization prediction, electronic health records, clustering, classification

1 Background and significance

Diabetes is recognized as the world's fastest growing chronic condition. One in 11 adults has diabetes worldwide (415 million) and 12% of global health expenditures is spent on diabetes ($673 billion).[1] In the USA alone, 29.1 million people or 9.3% of the population had diabetes in 2012.[2] Given its impact, medical and health services studies have been tracking the prevalence and trends in diabetes among adults.[3–5] While diabetes affects primarily the patients at many levels (physical, financial, etc.), it also poses an economic burden to states influencing healthcare costs and GDP/productivity metrics.

The U.S. healthcare system is undoubtedly expensive, excellent at treating acute conditions but ineffective at keeping patients out of the hospital.[6,7] Hospital care accounts for 31% of U.S. healthcare spending,[8] the latter

Center for Information and Systems Engineering, Boston University, Boston, MA, USA

Corresponding author:
Ioannis Ch. Paschalidis, Department of Electrical and Computer Engineering, Division of Systems Engineering, Boston University, 8 Saint Mary's St., Boston 02215, MA, USA.
Email: yannisp@bu.edu

totaling $3 trillion or 17% of GDP annually. A recent study, however, found that nearly $30.8 billion in hospital costs in the year 2006 were potentially avoidable,[9] with diabetes-related hospitalizations accounting for 19% ($5.8 billion) of this amount. Consequently, even a modest percentage reduction in unnecessary hospitalizations, achieved by better controlling the disease in an outpatient setting, can result in sizeable savings. Prevention requires prediction and this motivates the work in this paper.

Two key enablers to such research are: the growing availability of patient Electronic Health Records (EHR) and the existence of sophisticated algorithms that can be learnt from the data. Surprisingly, not until recently have EHRs been used in conjunction with advanced algorithms,[10,11] even though they have been shown to lead to better care.[12] Predictive methods, in particular, have been used for example in the context of heart-related problems,[13–15] hemodialysis,[16] diabetes in older adults,[17–20] and multiple disease prediction.[21] To the best of our knowledge, predicting diabetes-related hospitalizations based on EHR history using machine learning algorithms is a novel problem.

Diabetes mellitus is a set of metabolic diseases affecting the body's ability to modulate blood sugar levels. Type I affects younger patients and is caused by the inability of the pancreas to produce enough insulin. Type II appears in older people when cells develop insensitivity to insulin. Gestational diabetes appears during pregnancy. Type II diabetes is by far the most common and in this paper we focus on patients with this type. Diabetes complications include nephropathy, neuropathy, retinopathy, vasculopathy (leading to heart disease and stroke), and foot ulcers. Many of these complications can lead to hospitalization; however, it is estimated that about 40% such hospitalizations do not list diabetes as a primary/secondary diagnosis.[22] To remove potential biases in the EHR, we will use a statistical method to associate different hospitalization types with diabetes.

2 Objective

We seek to predict hospitalizations associated with Type II diabetes within one year from the time the EHR of a patient is examined. We will treat hospitalization prediction as a classification problem, distinguishing between patients likely to be hospitalized or not. Intuitively, however, patients belong to different clusters depending on their demographics and ailments that are likely to cause a future hospitalization. Common supervised learning methods can certainly make classifications without considering these hidden clusters; yet, identifying the clusters can potentially improve classification performance. More importantly, hidden cluster identification yields results that are easier to *interpret*.

Patients in the same cluster, especially if the cluster is identified based on a low-dimensional subspace of "diagnostic" features, share key characteristics (including potentially race and ethnicity) and their cluster membership offers an explanation as to why they have been flagged for a future hospitalization. In the medical setting, *interpretability* has an essential role in persuading physicians to trust the learning outputs and rely on them for their decision making. EHRs exhibit interesting special structure in that for each patient only a very low-dimensional subset of features is important in predicting a future hospitalization. This subset is different for each cluster and, typically, there is no universal set of irrelevant features that can be eliminated.[13,23] This suggests that it is useful to consider sparse classifiers for each cluster. Sparse models have gained popularity in the literature for their interpretability and superior (out-of-sample) performance.[24,25]

The remainder of the paper is organized as follows. In Section 3, we discuss methods we apply to the hospitalization prediction problem. We propose a novel method, an alternating optimization approach, which jointly discovers the clusters in the class of hospitalized patients and optimizes the classifiers that separate each cluster of hospitalized patients from the non-hospitalized patients. We establish the convergence of this joint clustering/classification process and characterize its Vapnik-Chervonenkis (VC) dimension[26] – a metric of complexity of the classification function that can lead to generalization guarantees. In Section 4, we describe the dataset used in our experiments and in Section 5 we present our experimental results. Conclusions are in Section 6.

3 Methods

We formulate the hospitalization prediction problem as a binary supervised classification problem. For each patient, we derive features from the EHR and we seek to differentiate between patients who will be hospitalized in a fixed target year (positive class) and patients who will not be admitted to the hospital in the target year (negative class). During the training of each classification model, both the features and the labels of the

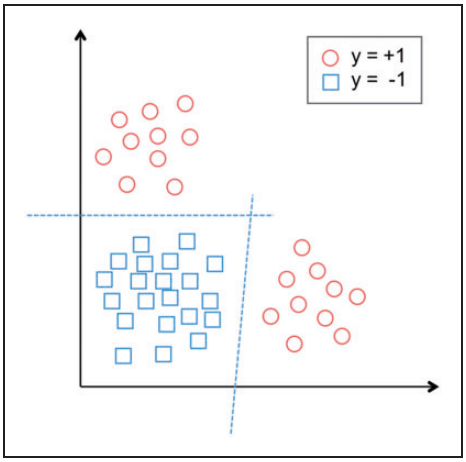

Figure 1. An example with two clusters in the positive class (red circles) separated by two different classifiers from the negative class (blue squares).

training set patients are known to the algorithm. We explore a variety of learning methods, such as support vector machines (SVMs) with various kernels,[27] random forests,[28,29] and the more computationally expensive gradient tree boosting.[29,30] We also experiment with sparse (l_1-regularized) versions of some algorithms; specifically, sparse SVMs and sparse logistic regression.

3.1 Our alternating clustering and classification framework

To develop this new framework, we consider a classification problem that has multiple hidden clusters in the positive class, while the negative class is assumed to be drawn from a single distribution. For different clusters in the positive class, we assume that the discriminative dimensions, with respect to the negative class, are different and sparse. We could think of these clusters as "local opponents" to the whole negative set (see Figure 1) and therefore, the "local boundary" (classifier) could naturally be assumed to be different and lying in a lower-dimensional subspace of the feature vector.

We propose a joint cluster detection and classification problem under the SVM framework. Let $(\boldsymbol{x}_i^+, y_i^+), i = 1, \ldots, N^+$, denote the $(D+1)-$dimensional positive samples, where $\boldsymbol{x}_i^+$ is the $D-$dimensional feature vector of sample i and $y_i^+ = 1$ the class label. Similarly, $(\boldsymbol{x}_j^-, y_j^-), j = 1, \ldots, N^-$, denote the negative samples with $y_i^- = -1$. Assuming L hidden clusters in the positive class, we wish to discover the L clusters (denoted by a mapping function $l(i) : \{1, \ldots, N^+\} \rightarrow \{1, \ldots, L\}$) and L sparse linear SVM classifiers, one for each cluster. Let $(\boldsymbol{\beta^l}, \beta_0^l)$ be the vector orthogonal to the SVM hyperplane for cluster l. Let also T^l be a parameter controlling local per-cluster sparsity. The joint problem is

$$
\begin{aligned}
&min_{\boldsymbol{\beta^l}, \beta_0^l, l(i)} \sum_{l=1}^{L} \left(\frac{1}{2} \|\boldsymbol{\beta}^l\|^2 + \lambda^+ \sum_{\{i:l(i)=l\}} \xi_i^l + \lambda^- \sum_{j=1}^{N^-} \zeta_j^l \right) \\
&s.t. \sum_{d=1}^{D} |\beta_d^l| \leq T^l, \qquad \forall l = 1, \ldots, L, \\
&\xi_i^{l(i)} \geq 1 - y_i^+ \beta_0^{l(i)} - \sum_{d=1}^{D} y_i^+ \beta_d^{l(i)} x_{i,d}^+, \qquad \forall i = 1, \ldots, N^+, \\
&\zeta_j^l \geq 1 - y_j^- \beta_0^l - \sum_{d=1}^{D} y_j^- \beta_d^l x_{j,d}^-, \qquad \forall l = 1, \ldots, L, \quad j = 1, \ldots, N^-, \\
&\xi_i^{l(i)}, \zeta_j^l \geq 0, \quad \forall i = 1, \ldots, N^+, \qquad j = 1, \ldots, N^-
\end{aligned}
\tag{1}
$$

The negative samples are not clustered but simply copied into each cluster. So their empirical costs are counted L times as shown in equation (1). The relative weight of costs from negative samples compared to that of the

positive samples is controlled by λ^- and λ^+. The constraint $\sum_{d=1}^{D}|\beta_d^l| \leq T^l$ is an l_1-relaxation of the sparsity requirement for the local classifiers.

Problem (1) involves two sets of decision variables: $(\boldsymbol{\beta}^l, \beta_0^l)$ for the classifiers and $l(i)$ for cluster assignment. As we have shown in Xu et al.,[15] the problem is a mixed integer programming problem, but given $l(i)$ it reduces to L quadratic optimization problems. This motivates the alternating optimization approach we present next. Preliminary work on such a method was reported in Xu et al.[15] for the problem of predicting hospitalizations due to heart diseases. The approach contains two major modules: (i) training a classifier for each cluster and (ii) re-clustering samples given all the estimated classifiers using a subset of "diagnostic" features $\mathcal{C}$. Note that since only positive samples belong to different clusters, only these samples need to be re-clustered. During the training phase, we alternate between (i) training L sparse classifiers and (ii) re-clustering the positive samples given the classifiers – until convergence. The algorithm for training and testing in the alternating clustering and classification (ACC) framework is shown in Algorithms 1 to 3. Algorithm 1 describes the training process, while Algorithm 2 provides details on how we re-cluster the positive samples given the classifiers learnt in (i). We note that training a sparse linear SVM classifier amounts to solving a quadratic programming problem, which can be done efficiently (in polynomial time to the size of the input, which is linear in the number of sample points in a cluster and the number of features D).

Algorithm 1. Alternating clustering and classification training.

Initialization:

For $i = 1, \ldots, N^+$, assign positive class sample i to cluster $l(i) \in \{1, \ldots, L\}$ (e.g. randomly).

repeat

Classification Step:

Train a sparse linear SVM classifier for each cluster. Each classifier is the outcome of a quadratic optimization (similar to equation (1) but specific to a single cluster) providing $(\boldsymbol{\beta}^l, \beta_0^l)$ and an optimal objective value O^l.

Re-clustering Step:

Re-cluster the positive samples using Alg. 2 and update the assignments $l(i)$.

until no $l(i)$ is changed or $\sum_l O^l$ does not decrease.

Algorithm 2. Re-clustering procedure given classifiers.

Input: positive samples $\boldsymbol{x}_i^+$, classifiers $(\boldsymbol{\beta}^l, \beta_0^l)$, current clusters assigning i to cluster $l(i)$.

For $i \in \{1, \ldots, N^+\}$ do

for all $l \in \{1, \ldots, L\}$ do

calculate the projection $a_i^l = \boldsymbol{x}_{i,C}^{'}\boldsymbol{\beta}_C^l$ of positive sample i onto the classifier for cluster l in the feature subspace corresponding to $\mathcal{C} \subseteq \{1, \ldots, D\}$.

end for

update the cluster assignment of sample i from $l(i)$ to $l^*(i) = \arg max_l a_i^l$, subject to

$$\boldsymbol{x}_i^{+'}\boldsymbol{\beta}^{l^*(i)} + \beta_0^{l^*(i)} \geq \boldsymbol{x}_i^{+'}\boldsymbol{\beta}^{l(i)} + \beta_0^{l(i)}. \quad (2)$$

end for

Algorithm 3. Classifying a new sample.

Input: A new (test) sample $\boldsymbol{x}$.

Assign $\boldsymbol{x}$ to cluster $l^* = \arg max_l \, x_C^{'}\boldsymbol{\beta}_C^l$.

Classify $\boldsymbol{x}$ with the classifier $(\boldsymbol{\beta}^{l^*}, \beta_0^{l^*})$ by $\text{sgn}(\boldsymbol{x}'\boldsymbol{\beta}^{l^*} + \beta_0^{l^*})$.

Once training has been performed with Algorithm 1, we can classify a newly presented sample not seen during training using Algorithm 3. Specifically, we compute the projections on each classifier and assign the new sample to the cluster with the largest projection value. We use the classifier of this cluster to classify the samples in the corresponding cluster. We note that tuning λ^+ and λ^- in ACC should be done globally, i.e. λ^+ and λ^- should be fixed across all clusters to guarantee convergence.

Theorem I establishes the convergence of ACC, while Theorem II characterizes its Vapnik-Chervonenkis (VC) dimension[26] and provides theoretical generalization guarantees. Intuitively, if we adopt a more complex model to fit a set of training samples, the model is more likely to be overfitting and has a lower chance to generalize well to the test samples. The VC-dimension offers a theoretical way of quantifying the complexity of a model and leads to a relationship between the training and the test error rate. Linear classifiers in the D dimensional space have VC-dimension $D + 1$.[26] The proofs of the two theorems are presented in Appendices A and B, respectively.

Let H denote the family of clustering/classification functions produced by ACC, where D denotes the maximum number of features used by a classifier. Let $R_N(h)$ (or simply R_N) denote the training error rate of classifier h on N training samples randomly drawn from an underlying distribution P. Let $R(h)$ (or simply R) denote the expected test error of h with respect to P.

Theorem I. For any $\complement$, the ACC process converges.

Theorem II. The VC-dimension of the class H is bounded by $V_{ACC} = (L+1)L(D+1)\log(\frac{e(L+1)L}{2})$. Then for any $\rho \in (0,1)$, with probability at least $1-\rho$, $R \leq R_N + 2\sqrt{2\frac{V_{ACC}\log\frac{2eN}{V_{ACC}}+\log\frac{2}{\rho}}{N}}$

Theorem I guarantees that for any choice of the subset $\complement$ of 'informative features' used for the re-clustering of the positive samples in Algorithm 2, the ACC process converges. We note that the joint problem (1) is non-convex, which suggests that a globally optimal solution is hard to obtain. ACC is a local optimization method and it is only guaranteed to converge to a local optimal solution. Strategies such as multi-start (i.e. starting from several initial points, using ACC until convergence, and retaining the best local minimum found) can be adopted to find a ''deep'' local optimum. The joint problem (1) can also be formulated as a maximization of a log-likelihood function (using the hinge-loss formulation of the SVM classifiers and raising it to an exponent), where the cluster membership can be captured by latent indicator variables. In such a formulation, ACC can be seen as a particular implementation of the EM algorithm,[31] where the E-step corresponds to cluster assignment and the M-step corresponds to obtaining classifiers for each cluster. It is possible to accelerate the speed of convergence by reducing the amount of work in the E-step, which is linear in the number of training samples. One possible approach has been used in Thiesson et al.[32]; specifically, rather than assigning all training samples to a cluster in each iteration, select a subset of samples, assign them to cluster, and then perform the per cluster classification steps. In our setting, $N = N^+ + N^-$ is large (on the order of tens of thousands) but not extremely large and we did not find it necessary to test such an approach.

Theorem II states that the out-of-sample error is close to the training error with high probability, which provides a rigorous generalization guarantee of our method.

3.2 Performance evaluation

To measure accuracy, the dataset is split into a training and a test set. The classification models are then trained using the features and the labels of the patients in the training set. In the testing phase, the models, given the patient's features, predict the corresponding label, which can be directly compared to the ground truth label. We report two error metrics: the *detection rate* (also referred to as *sensitivity*), which measures how many patients out of the truly hospitalized patients were predicted to be hospitalized, and the *false alarm rate* (equals one minus the so called *specificity*), which measures how many patients out of the truly non-hospitalized were predicted to be hospitalized.[a] Two other related metrics are the *precision*, defined as the ratio of the number of the truly hospitalized over the number of the predicted to be hospitalized, and the *recall*, which is the same as the detection rate or sensitivity. Many pairs of these error metrics can be generated by changing the decision threshold in the classification models. We plot the detection rate versus the false alarm rate in the receiver operating characteristic (ROC) curve and the precision versus the recall in the precision-recall curve (PRC). Typically, one chooses a point on the ROC (or PRC) to operate on, depending on the application, i.e. how high false alarm rate or low detection rate one can afford. The Area Under the ROC Curve (AUC) and the Area Under the Precision-Recall Curve (AUC-PR) are summary statistics, taking values between 0 and 1, that allow us to compare different ROC and PRC curves, respectively. The higher the AUC (AUC-PR) value of the ROC (PRC) curve, the better. A completely random assignment of patients into the two classes (hospitalized and non-hospitalized) has an AUC of 0.5 and an AUC-PR equal to the proportion of positive samples (precision is constant despite of the changing recall in the PRC). It is worth noting that unlike the ROC curve, the PRC is not guaranteed to be monotonic.[33]

While ROC curves could provide misleading interpretation of specificity when utilized in imbalanced classification cases,[34,35] (i.e., when one class has many more samples than the other), the PRC presents a more accurate measurement of imbalanced classification performance because it also considers the fraction of true positive samples among all positive predictions.[34,35] In the diabetes-related hospitalization prediction problem under consideration, this imbalance is present as there are many more non-hospitalized patients than hospitalized.

Table 1. Medical factors.

Ontology	Examples
Demographics	Sex, age, race
Diagnoses	For example, Diabetes mellitus with complications, Thyroid disorders, Hypertensive disease, Pulmonary heart disease, Heart failure, Aneurysm, Skin infections, Abnormal glucose tolerance test, Family history of diabetes mellitus
Procedures (CPT or ICD9)	For example, Procedure on single vessel, Insertion of intraocular lens prosthesis at time of cataract extraction, Venous catheterization, Hemodialysis, Transfusion of packed cells
Admissions	For example, Diabetes (with and without) complications, Heart failure and shock, Deep Vein Thrombophlebitis, Renal failure, Chest pain, Chronic obstructive pulmonary disease, Nutritional & miscellaneous metabolic disorders, Bone Diseases & Arthropathies, Kidney & urinary tract infections, Acute myocardial infarction, O.R. procedures for obesity, Hypertension
Laboratory Test Values	Hematology, Chemistry, Urinalysis, Coagulation tests
Vital Signs	For example, Blood pressure, Pulse, Respiratory rate, Temperature, Body Mass Index (BMI)
Blood Glucose Regulation Agents	Insulin, Anti-hypoglycemic, Oral hypoglycemic agents, etc.
Service by department	Inpatient (admit), Inpatient (observe), Outpatient, Emergency Room

Consequently, we present both ROC (AUC) and PRC (AUC-PR) curves to enable a comprehensive understanding of the classification performance.

As we commented earlier, the interpretability of the results is critical in ensuring practical use. We will assess interpretability in terms of highlighted important features that are the most helpful into making the classification decision. In the ACC approach, the discovered clusters bear a lot of information as to why the hospitalization occurred. To visualize this information, we listed the most distinguishable mean feature values over the patients in each cluster, which correspond to key features shared by the patients in the cluster.

4 Materials: the diabetes dataset

The data used for the experiments come from Boston Medical Center (BMC). BMC is the largest safety-net hospital in New England and with 13 affiliated Community Health Centers (CHCs) provides care for about 30% of Boston residents. The population of the study consists of patients with at least one diagnosis record of diabetes mellitus (ICD9 code 250) between 1 January 2007 and 31 December 2012. For each patient in the above set, we extract the medical history (demographics, visit history, problems, procedures and department information) during the period 1 January 2001 to 31 December 2012. The data we process for these patients come from the hospital EHR and billing systems, which record admissions or visits and the primary diagnosis/reason. The diabetes-related medical history of the patients is described by various categories of medical factors (that we identified using feedback from doctors), which, along with some examples corresponding to each, are shown in Table 1. As expected, many of the diagnoses and procedures are direct complications due to diabetes. Diabetes-related admissions are not trivially identifiable, and are revealed through the procedure described in the next section. Overall, this dataset consists of 40,921 patients.

In more detail, with every patient visit to the hospital, at least one record with a medical factor and a time-stamp containing the admittance date (and the discharge date) is created. In order to organize all the information available in some uniform way for all patients, some pre-processing of the data is required. Details will be discussed in a later section. We will refer to the summarized information of the medical factors over a specific time interval as features. Each feature related to Diagnoses, Procedures CPT (Current Procedural Terminology), and Procedures ICD9 (International Classification of Diseases, 9th edition) and visits to each Department is an integer count of such records for a specific patient during the specific time interval. Zero indicates the absence of any record.

4.1 Identifying the diabetes-related hospitalizations/admissions

Identifying the hospitalizations that could be attributed to diabetes is not a trivial task, because for financial reasons (i.e. higher reimbursement), many diabetes-related hospitalizations are recorded in the system as other

types of admissions, e.g. heart-related. To that end, we conduct a complementary statistical study to determine which types of admissions are diabetes-related. For simplicity, we adopt the classic p-value approach. There are also alternative but more complex hypothesis testing methods, including the critical value test[36] and methods involving confidence intervals.[37]

We consider all patients with at least one admission record between 01 January 2007 and 31 December 2012. From this set, patients with at least one diabetes mellitus record are assigned to the diabetic population, while the rest are assigned to the non-diabetic population. We list the union of all the unique admission types for both populations (732 unique types). The total number of admissions for the diabetic and non-diabetic populations is $N_1 = 47,452$ and $N_2 = 116,934$, correspondingly. For each type of admission d, each admission event can be represented as a binary random variable that takes the value 1, if the hospitalization occurs because of this type of admission, or 0 otherwise. Thus, we can transform the two sets of admission records for the two populations into 0/1 sequences. By (statistically) comparing the proportions of d in the two populations, we can infer whether admission d was caused mainly by diabetes or not.

At this point, we will elaborate on a statistical hypothesis test that involves sample differences of proportions.[38] Suppose we generate two sets of admissions of size N_1 and N_2 drawn from the diabetic and the non-diabetic patient populations, respectively. Consider a specific admission type d and suppose that it appears with probability p_1, out of all possible admission types, in the diabetic population. Similarly, a type d admission appears with probability p_2 in the non-diabetic population. Given now the two sets of admissions from diabetics and non-diabetics of size N_1 and N_2, let P_1 and P_2 be the corresponding sample proportions of type d admissions. We want to statistically compare P_1 and P_2 and assess whether a type d admission is more prevalent in the diabetics vs. the non-diabetics. Consider as the null hypothesis the case where $p_1 = p_2$, i.e. a type d admission is equally likely in the two populations. Under the null hypothesis, the sampling distribution of differences in proportions is approximately normally distributed, with its mean and standard deviation given by $\mu_{P_1-P_2} = 0$ and $\sigma_{P_1-P_2} = \sqrt{pq(\frac{1}{N_1} + \frac{1}{N_2})}$, where $p = \frac{N_1P_1+N_2P_2}{N_1+N_2}$ is used as an estimate of the probability of a type d admission in both populations and $q = 1 - p$. By using the standardized variable $z = \frac{P_1-P_2}{\sigma_{P_1-P_2}}$, we can assess if the results observed in the samples differ markedly from the results expected under the null hypothesis. We do that using the single sided p-value of the statistic z. The smaller the p-value is, the higher the confidence we have in the alternative hypothesis or, equivalently, in the fact that the diabetic patients have higher chance of getting admission records of type d than the non-diabetic ones (since we consider the difference $P_1 - P_2$). We list admission types in increasing order of p-value and we set a threshold of p-value $\leq \alpha = 1E-4$; admission types with p-value less than α are considered to be attributed to diabetes.

4.2 Data pre-processing

The features are formed as combinations of different medical factors (instead of considering the factors as separate features) that better describe what happened to the patients during their visits to the hospital. Specifically, we formulate triplets that consist of a diagnosis, a procedure (or the information that no procedure was done) and the service department. An example of a complex feature (a triplet) is the diagnosis of ischemic heart disease that led to an adjunct vascular system procedure (procedure on a single vessel) while the patient was admitted to the inpatient care. Clearly, since each category can take one of several discrete values, a huge number of combinations should be considered. Naturally, not all possible combinations occur, which reduces significantly the total number of potential features that describe each patient. Also for each patient, we extract information about the diabetes type over their history (keeping only patients with Type II) and demographics including age, gender and race.

Next, we present several data organization and pre-processing steps we take. For each patient, a target year is fixed and all past patient records are organized as follows.

- *Setting the target time interval to a calendar year*. Based on some preliminary experiments we conducted, we observed that there is greater variability in the results when trying to predict hospitalizations in periods of time shorter than a year (e.g. predicting hospitalization in the next one, three or six months). Thus, we have designed our experiment to predict hospitalizations in the target time interval of a year starting on the 1 January and ending on 31 December.[b]
- *Selection of the target year*. As a result of the nature of the data, the two classes are highly imbalanced. To increase the number of hospitalized patient examples, if a patient had only one hospitalization throughout 2007–2012, the year of hospitalization will be set as the target year. If a patient had multiple hospitalizations,

a target year between the first and the last hospitalizations will be randomly selected. The year 2012 is set as the target year for patients with no hospitalization, so that there is as much available history for them as possible.

- *Removing patients with no record*. Patients who have no records before the target year are removed, since there is nothing on which a prediction can be based. The total number of patients left is 33,122, including the 26,478 patients with Type II diabetes under consideration. After this process, the proportion of hospitalized patients with Type II diabetes in the dataset is 13.48% (3570 out of 26,478).
- *Forming the complex features*. We create a diagnoses-procedures-service department indicator triplet (complex feature) to keep track of which diagnosis occurs with which procedure and service department. The procedures that are not associated with any diabetes-related diagnosis are removed. Diagnoses in the dataset are listed in the most detailed level of the ICD9 coding system. We group together procedures that belong to the same ICD/CPT family, resulting in 31 categories (out of 2004 in total).
- *Summarization of the complex features in the history of a patient*. We form four time blocks for each medical factor. Time blocks 1, 2, and 3 summarize the medical factors over one, two, and three years before the target year, whereas the fourth time block averages all earlier records. Naturally not all combinations of diagnoses-procedures-service department occur, and we only keep the triplets that occur; then adding the demographic features produces a 9402-dimensional vector of features characterizing each patient.
- *Reducing the number of complex features*. We remove all the features that do not contain enough information for a significant amount of the population (less than 1% of the patients), as they could not help us generalize. This leaves 320 complex medical and three demographic features.
- *Other detailed information*. We also consider 245 more detailed medical features, including lab test values, vital signs and blood glucose regulation agents (see Table 1). By calculating the average lab test values, average vital signs or existence of regulation agents in the four time blocks, we obtain $245 \times 4 = 980$ additional features. Removing features with standard deviation less than $1E-4$ reduces the number of features to 945. Together with the features we described earlier, this results in $945 + 3 + 320 = 1268$ features.
- *Identifying the diabetes type*. The ICD9 code for diabetes is assigned to category 250 (diabetes mellitus). The fifth digit of the diagnosis code determines the type of diabetes and whether it is uncontrolled or not stated as uncontrolled. Keeping only the 26,478 patients with Type II diabetes, we have two types of diabetes diagnoses: Type II, not stated as uncontrolled (fifth digit 0), and Type II or unspecified type, uncontrolled (fifth digit 2). Based on these types, we count how many records of each type each patient had in the four time blocks before the target year, thus adding eight new features for each patient.
- *Splitting the data into a training set and a test set randomly*. As is common in supervised machine learning, the population is randomly split into a training and a test set. Since from a statistical point of view all the data points (patients' features) are drawn from the same distribution, we do not differentiate between patients whose records appear earlier in time than others with later time stamps. A retrospective/prospective approach appears more often in the medical literature and is more relevant in a clinical trial setting, rather than in our algorithmic approach. What matters in our setting is that for each patient prediction we make (hospitalization/non-hospitalization in a target year), we only use that patient's information before the target year.

5 Experimental results and discussion

5.1 Performance evaluation

We evaluate classification performance out-of-sample, i.e. in a test set not seen during training. Figures 2 and 3 plot ROC and PRC curves for a variety of classification methods, respectively; Tables 2 and 3 list the corresponding AUCs and AUC-PRs (average and standard deviation of AUC and AUC-PR over 10 runs with different training and test sets). Parameter tuning was done for all methods using k-fold cross validation. For ACC, the initial assignments of the positive samples to the clusters are obtained from k-means clustering, and multi-start is implemented to find the best local optimum. The parameters used in equation (1) are set as follows. The number of clusters L explicitly takes its values from $\{2, 3, 4\}$ for all methods involving clustering; the soft-margin parameter for the negative class λ^- takes its values from $\{100, 10, 1, 0.1\}$; and the soft-margin parameter for the positive class λ^+ is set equal to $L\lambda^-$. Some preliminary experiments led us to set the sparsity-controlling parameter $T' = 12$ to save on computational cost. For ACC, we employ one more innovation to improve the prediction results. Specifically, for each cluster, we solve several instances of the per-cluster sparse SVM as follows.

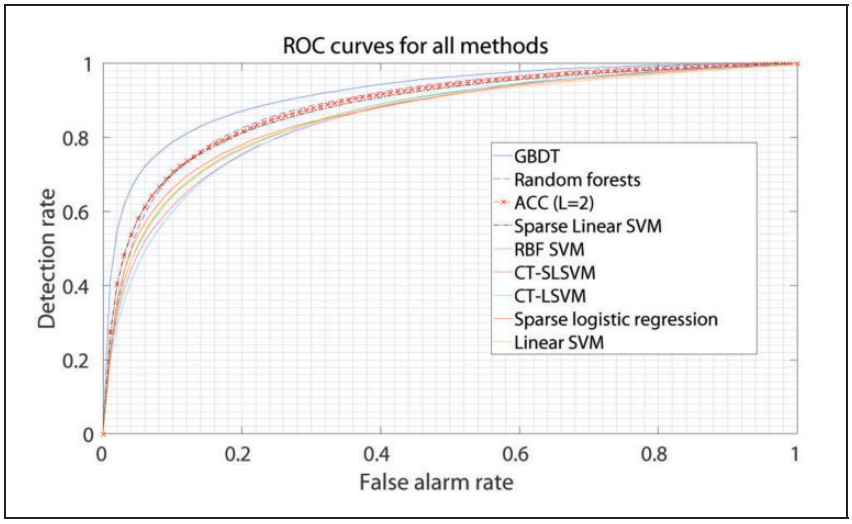

Figure 2. Receiver operating characteristic curves for various classification methods.

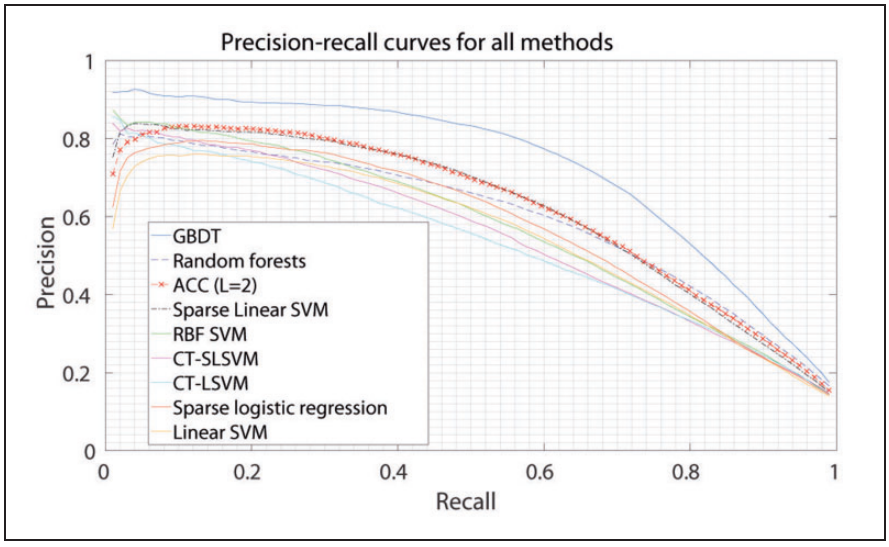

Figure 3. Precision-recall curves for various classification methods.

Table 2. Average (avg) and standard deviation (std) of the area under ROC curve (AUC) of various methods we have experimented with over 10 runs.

Method	Average AUC	Std AUC	Method	Average AUC	Std AUC
ACC, $L = 1$	0.8814	0.0025	Linear SVM	0.8531	0.0029
ACC, $L = 2$	0.8861	0.0032	RBF SVM	0.8594	0.0037
ACC, $L = 3$	0.8829	0.0039	sparse logistic regression	0.8613	0.0027
ACC, $L = 4$	0.8812	0.0027	Random Forests	0.8882	0.0054
CT-SLSVM ($L = 2$)	0.8522	0.0034	Gradient Tree Boosting (GBDT)	0.9190	0.0033
CT-LSVM ($L = 2$)	0.8502	0.0072			

ACC: Alternating clustering and classification.

Table 3. Average and standard deviation of the area under precision-recall curve (AUC-PR) of various methods we have experimented with over 10 runs.

Method	Avg AUC-PR	Std AUC-PR	Method	Avg AUC-PR	Std AUC-PR
ACC, $L = 1$	0.6200	0.0104	Linear SVM	0.5512	0.0129
ACC, $L = 2$	0.6214	0.0106	RBF SVM	0.5758	0.0110
ACC, $L = 3$	0.6085	0.0115	sparse logistic regression	0.5752	0.0091
ACC, $L = 4$	0.6035	0.0109	Random Forests	0.6003	0.0159
CT-SLSVM ($L = 2$)	0.5518	0.0124	Gradient Tree Boosting (GBDT)	0.7272	0.088
CT-LSVM ($L = 2$)	0.5355	0.0175			

ACC: Alternating clustering and classification.

First, we solve the problem with all features and a fixed T^l. This has the effect of selecting a subset of the features. Then, we solve a second instance of the problem using only the subset of the features selected. We keep iterating in this fashion until a relatively small subset of features is being used.

For all methods 40% of the data are used for training and the rest for testing. The training data are normalized to have zero mean and unit standard deviation and are balanced by down-sampling the negative population. We also compare ACC with two other hierarchical approaches that first cluster the data using the k-means clustering[39] and then perform the classification task using linear SVM (we denote the method as CT-LSVM) and sparse (l_1-regularized) linear SVM (we denote the method as CT-SLSVM). Only the best results for CT-LSVM (obtained under $L = 2$) and CT-SLSVM (obtained under $L = 2$) are presented.

Clustering with ACC can use a subset of "diagnostic" features (subset $\mathcal{C}$ in Algorithm 2), since these are the features that better delineate across different types of diabetes complications. We base, however, the clustering in our experiments on all features due to the fact that almost all triplet features are related to "diagnostic" features. The results indicate that Gradient Tree Boosting (GBDT) outperforms all other methods in terms of AUC in Table 2 and AUC-PR in Table 3. Random forest comes second in terms of AUC with ACC close behind, while ACC comes second in terms of AUC-PR with random forest third. Both GBDT and random forests produce a very complex classifier involving hundreds of decision trees. As such they lack interpretability, which, as we argued, is a critical consideration.

ACC, on the other hand, is able to detect the hidden positive clusters and identify why a specific patient is labeled as hospitalized. Among ACC variants, the best performance is obtained for $L = 2$ clusters, with the performance of the other variants using more clusters being close behind. The fact that ACC ($L = 2$) is better than ACC ($L = 1$) illustrates that appropriate clustering can not only produce meaningful cluster interpretations, but also improve classification performance compared to the base (SVM) classifier used in each cluster. It is interesting that ACC performs quite well even though the resulting classifiers are relatively sparse and do not use many features. This also makes them easy to implement. Notice that ACC utilizes sparse linear SVM as the base classifier. According to Theorem II, sparsity (i.e. small D) leads to smaller generalization error. ACC also proved to be efficient from a computational point of view, since in our implementation, it is faster than random forests by a factor of 3, and faster than Gradient Tree Boosting (GBDT) by a factor of 5.3. Figure 4 shows how the objective function decreases during ACC iterations.

In an attempt to interpret the ACC clusters, we list in Table 4 the mean value over each cluster of the features used by the per-cluster classifiers. This is done for a single repetition of the experiment and $L = 2$, yielding interesting clusters and highlighting the interpretative power of ACC. We concentrate on the most distinguishable features in the clusters. Specifically, for each feature we used Welch's t-test to compute a two-tailed p-value, where the null hypothesis was that the two cohorts (patients in cluster 1 and cluster 2) have equal means. All the features listed in Table 4 have a p-value less than 1E-3, providing strong evidence against the null hypothesis. ACC assigns 51.87% of hospitalized patients in the training set to cluster 2 and the remaining to cluster 1. We observe that hospitalized patients in Cluster 1 are older, have more hypertension and heart failure (measured in avg. number of diagnoses), take more drugs for heart diseases (measured in average number of drugs taken), and have indicators of renal disease (higher serum creatinine values and higher blood urea nitrogen). Hospitalized patients in Cluster 2 have diabetes with not as significant heart disease complications, indicated as diabetes with no associated procedures, hospitalizations, and, in general, not stated as uncontrolled. This is quite interesting and consistent with earlier work that identified the relationship between diabetes and

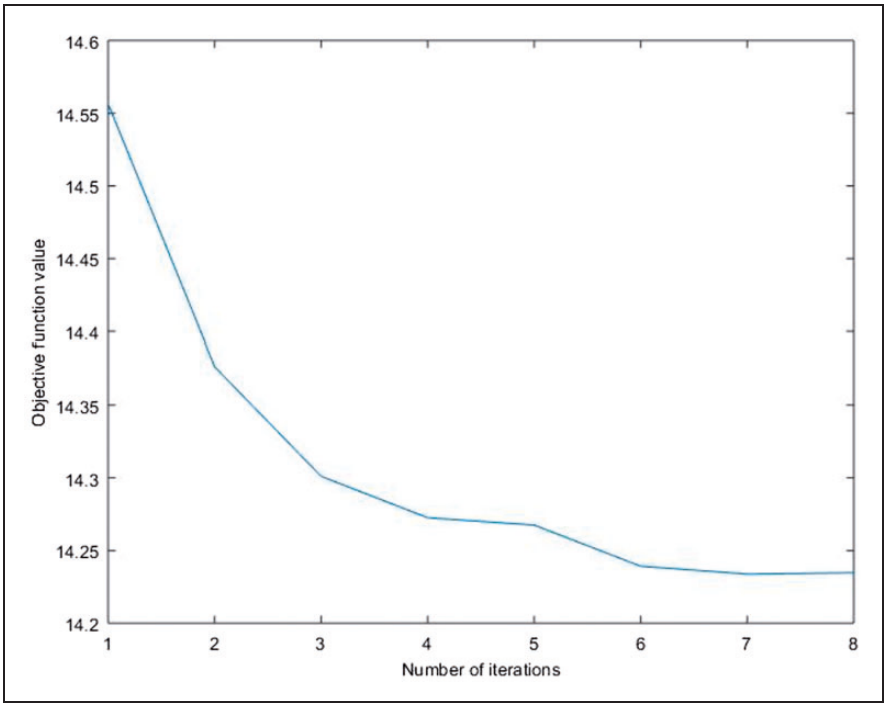

Figure 4. Objective function value as a function of alternating clustering and classification iterations.

Table 4. Average feature values in the clusters produced by ACC ($L = 2$).

Variables	Mean in Cluster 1	Mean in Cluster 2	p-value	Variables	Mean in Cluster 1	Mean in Cluster 2	p-value
Age	72.98	66.11	1.04E-18	Creatinine, serum	1.27	0.96	7.85E-11
Hypertensive diseases (diagnoses)	1.52	1.16	5.44E-04	Glucose, point of care	148.65	162.23	7.33E-38
Heart failure (diagnoses)	0.31	0.09	5.69E-09	Platelet count	226.40	263.63	1.67E-44
Diabetes mellitus, no procedure emergency room	0.13	0.23	6.57E-07	Review/Order Lab tests	0.43	0.25	2.70E-19
Diabetes mellitus, no procedure in patient (Observe)	0.94	1.48	2.92E-10	Review/Order Radiology	0.23	0.14	3.87E-08
Diabetes Type II, not stated as uncontrolled	1.86	3.70	1.73E-25	Review/Order other tests	0.22	0.10	2.34E-17
Cardiology-related medicine	0.75	0.18	3.12E-16	Urea nitrogen, blood	21.77	16.80	3.42E-26

ACC: alternating clustering and classification.

specific complications (heart disease in our case).[18,40] It appears that the method is identifying a cluster of patients with diabetes and heart disease and is using a different classifier for these patients compared to the remaining patients.

We note that the *p*-values in Table 4 are reported without adjustment for multiple hypothesis testing, i.e. each *p*-value corresponds to the probability that the cluster means of each variable in isolation are as reported under the null hypothesis. In order to control the false discovery rate of the multiple hypothesis test, we applied the Benjamini–Yekutieli procedure to adjust the *p*-values.[41,42] All of the adjusted *p*-values are below 1E-6 except for two variables: ''Hypertensive Diseases (diagnoses)'' and ''Diabetes Mellitus No Procedure Emergency Room'', which have adjusted *p*-values equal to 1.77E-03 and 2.30E-06, respectively. These can be seen as sufficiently small to provide strong evidence for rejecting the null hypothesis.

5.2 Cost–benefit analysis

We next assess the potential financial benefits of using a predictive model like the one we developed. We take year 2012 as an example; our dataset has $N_H = 619$ hospitalized and $N_{NH} = 22{,}616$ non-hospitalized patients that year. According to Clancy et al.,[43] the average cost per hospitalization due to diabetes with complications is \$9,500. Thus, assuming no spending on the non-hospitalized patients and a single hospitalization for the hospitalized, the expected cost per patient if no prevention measures are implemented is

$$\frac{9500N_H}{N_H + N_{NH}} = \$253 \tag{3}$$

Suppose now we elect to utilize the predictive model and operate at a point on the ROC curve corresponding to a roughly $P_D = 81\%$ detection rate and a $P_{FA} = 20\%$ false alarm rate (see Figure 2). We bring each patient predicted to be hospitalized to the clinic, at a cost of \$220 for a visit according to Clancy et al.,[43] and prescribe an one-year supply of drugs at an average cost of \$100. Additional recommendations involving lifestyle changes and social support may also be offered. Accounting for only the cost of the visit and the drugs, the cost of preventive measures is \$320 per patient. Notice that this overestimates the cost because for some patients predicted to be hospitalized, the physician may decide that additional drugs are not needed. For patients the predictive model misses, there is no action and they would receive their normal care. Let P_S be the probability that prevention is effective and averts the hospitalization. It follows that the cost per patient becomes

$$\frac{9500N_H(1 - P_D) + 320N_{NH}P_{FA} + 320N_HP_DP_S + (9500 + 320)N_HP_D(1 - P_S)}{N_H + N_{NH}} \tag{4}$$

A simple calculation implies that the above quantity is less than \$253 for $P_S > 0.34$. Taking $P_S = 0.5$ leads to an expected cost per patient equal to \$219, resulting in savings of \$34 per patient. If such a model was used for each patient with diabetes in the U.S. during 2002 (29.1 million), the overall savings amount to about \$1 billion for the year! This is about 17% of the overall amount spent on preventable diabetes-related hospitalizations each year.

5.3 Limitations of our study

While the results we have presented seem very promising and we have provided theoretical guarantees about the generalization ability of our proposed methodology, our study naturally suffers from data limitations. This is because we focus on patients from a specific hospital, coming mainly from lower socioeconomic classes. Our data come exclusively from the BMC system and we do not have access to any data corresponding to visits or treatment outside BMC. Specifically, patients our methods find to be at a high risk of hospitalization based on the available data, may have received treatment elsewhere which reduced their hospitalization risk, eventually avoiding a hospitalization we may have predicted. Moreover, patients who we predict will not be hospitalized may have been seen outside BMC and data captured for them may explain a future hospitalization. In both cases, in the absence of data about a patient, the predictive model is powerless. We conjecture, however, that the effect of such lack of data does not substantially alter the metrics on the predictive power of our methods. The main reason is that BMC patients are typically seen within the system because they lack the financial resources to receive care elsewhere. In any case, if additional data (e.g. from insurance claims) become available, they can be readily used by our methods to improve the predictions.

6 Conclusions

Diabetes is the fastest growing chronic condition causing a number of preventable hospitalizations. Diabetes is also associated with serious complications, such as heart disease and stroke, retinopathy, kidney failure, and lower-limb amputation. Early detection and treatment can slow down the progression of the disease and result in better health outcomes and huge savings. We considered the problem of predicting diabetes-related hospitalizations using information in the Electronic Health Records of the patients. We introduced a statistical procedure to identify the diabetes-related admissions and we experimented with a number of machine learning methods that predict hospitalizations in a target year for diabetic patients. With a 20% false alarm rate, we can correctly predict about 81% of the hospitalized patients while providing insight as to why each prediction is made. To that end, we developed a novel clustering and classification framework (ACC) that jointly discriminates between hospitalized and non-hospitalized patients and

discovers clusters of patients with key factors, different in each cluster, that lead to hospitalization. The identification of the clusters has the significant advantage of interpretability, which is crucial in the medical domain. We proved convergence of the new algorithm and established theoretical generalization guarantees. The proposed algorithm has wider applicability and the potential to be applied to other medical case studies, helping, for example, discover cohorts of patients with similar underlying issues and devising cohort-specific predictive models.

Acknowledgements

The authors would like to thank William Adams for providing access to the data. We thank William Adams, Theodora Anagnostou, Theofanie Mela, and Venkatesh Saligrama for useful discussions. We also thank an anonymous reviewer whose comments have led us to improve the paper.

Authors' contributions

WD and TSB co-designed the methods; TSB, TX and TW performed the analysis, produced results and figures, co-wrote the manuscript. IChP led the study, co-designed the methods, and co-wrote the manuscript.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research is partially supported by the NSF under grants DMS-1664644, CNS-1645681, CCF-1527292, IIS-1237022, by the ARO under grant W911NF-12-1-0390, by the ONR under grant MURI N00014-16-1-2832, by the NIH under grant 1UL1TR001430 to the Clinical & Translational Science Institute at Boston University, and by the Boston University Digital Health Initiative and the Center for Information and Systems Engineering.

Notes

a. The terms *detection rate* and *false alarm rate* are commonly used in the machine learning community, while the terms *sensitivity* and *specificity* appear more often in the medical literature.
b. The fact that we have chosen calendar years and not a sliding time window is a design choice we have made.

ORCID iD

Taiyao Wang http://orcid.org/0000-0002-0331-3892
Ioannis Ch. Paschalidis http://orcid.org/0000-0002-3343-2913

References

1. International Diabetes Federation. *Diabetes Atlas*, http://www.diabetesatlas.org/component/attachments/?task=download&id=116 (2015).
2. Center for Disease Control and Prevention. *National diabetes statistics report: estimates of diabetes and its burden in the United States, 2014*. US Department of Health and Human Services, https://www.cdc.gov/diabetes/pdfs/data/2014-report-acknowledgments.pdf (2014).
3. Menke A, Casagrande S, Geiss L, et al. Prevalence of and trends in diabetes among adults in the United States, 1988–2012. *JAMA* 2015; **314**: 1021–1029.
4. King H, Aubert RE and Herman WH. Global burden of diabetes, 1995–2025: prevalence, numerical estimates, and projections. *Diab Care* 1998; **21**: 1414–1431.
5. Rathmann W and Giani G. Global prevalence of diabetes: estimates for the year 2000 and projections for 2030. *Diab Care* 2004; **27**: 2568–2569.
6. Kangovi S, Barg FK, Carter T, et al. Understanding why patients of low socioeconomic status prefer hospitals over ambulatory care. *Health Aff (Millwood)* 2013; **32**: 1196–1203.
7. Thygesen LC, Christiansen T, Garcia-Armesto S, et al. Potentially avoidable hospitalizations in five European countries in 2009 and time trends from 2002 to 2009 based on administrative data. *Eur J Public Health* 2015; **25**: 35–43.
8. Goldman D and McGlynn E. *US health care: Facts about cost, access, and quality*. RAND Corporation, http://www.rand.org/content/dam/rand/pubs/corporate_pubs/2005/RAND_CP484.1.pdf (2005).
9. Jiang HJ, Russo CA and Barrett ML. *Nationwide Frequency and Costs of Potentially Preventable Hospitalizations, 2006. HCUP Statistical Brief# 72*. Rockville, MD: US Agency for Healthcare Research and Quality, http://www.hcupus.ahrq.gov/reports/statbriefs/sb72.jsp (2010).

10. Wu J, Roy J and Stewart WF. Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. *Med Care* 2010; **48**: S106–S113.
11. Weller GB, Lovely J, Larson DW, et al. Leveraging electronic health records for predictive modeling of post-surgical complications. *Stat Meth Med Res* 2018; 27: 3271–3285.
12. Featherstone I and Keen J. Do integrated record systems lead to integrated services? An observational study of a multi-professional system in a diabetes service. *Int J Med Inf* 2012; **81**: 45–52.
13. Dai W, Brisimi TS, Adams WG, et al. Prediction of hospitalization due to heart diseases by supervised learning methods. *Int J Med Inf* 2015; **84**: 189–197.
14. Dai W, Brisimi TS, Xu, Tingting, et al. A joint clustering and classification approach for healthcare predictive analytics. In: *2nd Workshop on data mining for medical informatics (DMMI 2015)*, San Francisco, CA, November 2015.
15. Xu T, Brisimi TS, Wang T, et al. A joint sparse clustering and classification approach with applications to hospitalization prediction. In: *55th Conference on decision and control (CDC), 2016 IEEE*. IEEE, pp.4566–4571.
16. Yeh J-Y, Wu T-H and Tsao C-W. Using data mining techniques to predict hospitalization of hemodialysis patients. *Decis Support Syst* 2011; **50**: 439–448.
17. Rosenthal MJ, Fajardo M, Gilmore S, et al. Hospitalization and mortality of diabetes in older adults: a 3-year prospective study. *Diab Care* 1998; **21**: 231–235.
18. Young BA, Lin E, Von Korff M, et al. Diabetes complications severity index and risk of mortality, hospitalization, and healthcare utilization. *Am J Manag Care* 2008; **14**: 15.
19. Tutun S, Khanmohammadi S, He L, et al. A meta-heuristic LASSO Model for diabetic readmission prediction. In: *Proceedings of the 2016 industrial and systems engineering research conference (ISERC)*. Anaheim, CA, May, 2016.
20. Zheng T, Xie W, Xu L, et al. A machine learning-based framework to identify type 2 diabetes through electronic health records. *Int J Med Inf* 2017; **97**: 120–127.
21. Bayati M, Bhaskar S and Montanari A. Statistical analysis of a low cost method for multiple disease prediction. *Stat Meth Med Res* 2016; 0962280216680242.
22. *Diabetes in America*. National Institute of Diabetes and Digestive and Kidney Diseases, NIH, http://www.niddk.nih.gov/about-niddk/strategic-plans-reports/Pages/diabetes-america-2nd-edition.aspx (1995, accessed January 2017).
23. Brisimi TS, Xu T, Wang T, et al. Predicting chronic disease hospitalizations from electronic health records: an interpretable classification approach. *Proc IEEE* 2018; **106**: 690–707.
24. Ng, Andrew Y. Feature selection, L 1 vs. L 2 regularization, and rotational invariance. In: *Proceedings of the twenty-first international conference on Machine learning, July 2004*, p. 78. ACM.
25. Bondell HD and Reich BJ. Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. *Biometrics* 2008; **64**: 115–123.
26. Vapnik V. *The nature of statistical learning theory*. Springer Science & Business Media, 2013.
27. Cortes C and Vapnik V. Support-vector networks. *Mach Learn* 1995; **20**: 273–297.
28. Breiman L. Random forests. *Mach Learn* 2001; **45**: 5–32.
29. Hastie T, Tibshirani R and Friedman J. *The elements of statistical learning: data mining, inference and prediction*. Berlin: Springer, 2001Springer series in statistics.
30. Mason L, Baxter J, Bartlett PL, et al. Boosting algorithms as gradient descent. In *Advances in Neural Information Processing Systems 12*. 2000, pp. 512–518. MIT Press.
31. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *J Stat Soc. Ser B (Methodol)*. 1977; 39(1): 1–38.
32. Thiesson B, Meek C and Heckerman D. Accelerating EM for large databases. *Mach Learn* 2001; **45**: 279–299.
33. Lever J, Krzywinski M and Altman N. Points of significance: classification evaluation. *Nat Meth* 2016; 13: 603–604.
34. Saito T and Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS ONE* 2015; **10**: e0118432.
35. Davis J, Burnside E, Dutra I, et al. View learning for statistical relational learning: with an application to mammography. In: *Proceedings of the nineteenth international joint conference on artificial intelligence*. Professional Book Center, Edinburgh, UK, 30 July – 5 August 2005, pp.677–683.
36. Rouder JN, Speckman PL, Sun D, et al. Bayesian t tests for accepting and rejecting the null hypothesis. *Psychon Bull Rev* 2009; **16**: 225–237.
37. Gardner MJ and Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. *Br Med J Clin Res Ed* 1986; **292**: 746–750.
38. Sprinthall, RC. *Basic Statistical Analysis*, 7th Edition, 2003. Boston, MA: Allyn and Bacon.
39. Hartigan JA and Wong MA. Algorithm AS 136: A k-means clustering algorithm. *J R Stat Soc Ser C Appl Stat* 1979; **28**: 100–108.
40. Nathan DM. Long-term complications of diabetes mellitus. *N Engl J Med* 1993; **328**: 1676–1685.
41. Benjamini Y and Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *Ann Stat* 2001; **29**: 1165–1188.
42. Benjamini Y. Discovering the false discovery rate: false discovery rate. *J R Stat Soc Ser B Stat Methodol* 2010; **72**: 405–416.

43. Clancy C, Munier W, Crosson K, et al. *2010 National healthcare quality & disparities reports*. Agency for Healthcare Research and Quality, NIH, http://health-equity.pitt.edu/2650/ (2011).
44. Sontag ED. VC dimension of neural networks. *NATO ASI Series F Computer and Systems Sciences*. New York, NY: Springer, 1998; 168: 69–96.
45. Bousquet O, Boucheron S and Lugosi G. Introduction to statistical learning theory. *Advanced lectures on machine learning*. Berlin, Heidelberg: Springer, 2004, pp.169–207.

Appendix 1. Proof of Theorem 1

Theorem 1. *For any* $\mathbb{C}$, *the ACC process converges*.

Proof. At each alternating cycle, for each cluster l we train a Sparse Linear SVM (SPLSVM) with positive samples of that cluster combined with all negative samples. This produces an optimal value O^l and the corresponding classifier $(\boldsymbol{\beta}^l, \beta_0^l)$. Specifically, the formulation is

$$\begin{aligned} O^l = & \min_{\boldsymbol{\beta}^l, \beta_0^l} \frac{1}{2}\|\boldsymbol{\beta}^l\|^2 + \lambda^+ \sum_{i=1}^{N_l^+} \xi_i^l + \lambda^- \sum_{j=1}^{N^-} \zeta_j^l \\ & s.t. \sum_{d=1}^{D} |\boldsymbol{\beta}_d^l| \le T^l, \\ & \xi_i^l \ge 1 - y_i^+ \beta_0^l - \sum_{d=1}^{D} y_i^+ \beta_d^l x_{i,d}^+, \quad \forall i \in \{1, \ldots, N_l^+\}, \\ & \zeta_j^l \ge 1 - y_j^+ \beta_0^l - \sum_{d=1}^{D} y_j^- \beta_d^l x_{j,d}^-, \quad \forall j \in \{1, \ldots, N^-\}, \\ & \xi_i^l, \zeta_j^l \ge 0, \quad \forall i \in \{1, \ldots, N_l^+\}, \quad \forall j \in \{1, \ldots, N^-\} \end{aligned} \tag{5}$$

We use the sum of the optimal objective function values in equation (5) across different clusters to prove convergence. We have

$$Z = \sum_{l=1}^{L} O^l = \sum_{l=1}^{L} \left(\frac{1}{2}\|\boldsymbol{\beta}^l\|^2 + \lambda^- \sum_{j=1}^{N^-} \zeta_j^l) + \lambda^+ \sum_{i=1}^{N_l^+} \xi_i^{l(i)} \right.$$

where $l(i)$ maps sample i to cluster $l(i)$, $\sum_{l=1}^{L} N_l^+ = N^+$, and $\boldsymbol{\beta}^l, \beta_0^l, \zeta_j^l,$ and $\xi_i^{l(i)}$ are optimal solutions of (5) for each l. Now, let us consider the change of Z at each iteration of the ACC procedure.

First, we consider the re-clustering step given SLSVMs. During the re-clustering step, the classifier and slack variables for negative samples are not modified. Only the $\xi_i^{l(i)}$ get modified since the assignment functions $l(i)$ change. When we switch positive sample i from cluster $l(i)$ to $l^*(i)$, we can simply assign value $\xi_i^{l(i)}$ to $\xi_i^{l^*(i)}$. Therefore, the value of Z does not change during the re-clustering phase and takes the form

$$Z = \sum_{l=1}^{L} \left(\frac{1}{2}\|\boldsymbol{\beta}^l\|^2 + \lambda^+ \sum_{\{i:l(i)=l\}} \xi_i^l + \lambda^- \sum_{j=1}^{N^-} \zeta_j^l \right)$$

Next, given new cluster assignments, we re-train the local classifiers by resolving problem (5) for each cluster l. Notice that re-clustering was done subject to the constraint in equation (2) (see Algorithm 1). Since $y_i^+ = 1$, we have

$$\xi_i^{l(i)} \ge 1 - y_i^+ \beta_0^{l(i)} - \sum_{d=1}^{D} y_i^+ \beta_d^{l(i)} x_{i,d}^+ \ge 1 - y_i^+ \beta_0^{l^*(i)} - \sum_{d=1}^{D} y_i^+ \beta_d^{l^*(i)} x_{i,d}^+$$

The first inequality is due to $\xi_i^{l(i)}$ being feasible for equation (5). The second inequality is due to $y_i^+ = 1$ and equation (2) in Algorithm (1). Thus, by assigning $\xi_i^{l(i)}$ to $\xi_i^{l^*(i)}$, it follows that the $\xi_i^{l^*(i)}$ remain feasible for problem (5). Given that the remaining decision variables do not change, $(\boldsymbol{\beta}^l, \beta_0^l, \zeta_j^l, \xi_i^{l(i)}, \quad \forall i = 1, \ldots, N_l^+, \quad \forall j = 1, \ldots, N^-)$ forms a feasible solution of problem (5). This solution has a cost equal to O^l. Re-optimizing can produce an optimal value that is no worse. It follows that in every iteration of ACC, Z is monotonically non-increasing. Given that Z is bounded below by zero, we establish the convergence of ACC.

Appendix 2. Proof of Theorem II

Theorem II. *The VC-dimension of the class H is bounded by* $(L+1)L(D+1)\log\left(e\frac{(L+1)L}{2}\right)$. *Then for any* $\rho \in (0,1)$, *with probability at least* $1-\rho$, $R \leq R_N + 2\sqrt{2\frac{V_{ACC}\log\frac{2eN}{V_{ACC}} + \log\frac{2}{\rho}}{N}}$.

Proof. The proof is based on Lemma 2 of Sontag.[44] Given an assignment of positive sample i to cluster $l(i)$ define L clustering functions

$$g_{l(i)} = \begin{cases} 1, & \text{if } l(i) = l, \\ 0, & \text{otherwise} \end{cases}$$

Hence, sample i is assigned to cluster $argmax_l g_l(i)$. This can be viewed as the output of $(L-1)L/2$ comparisons between pairs of g_{l_1} and g_{l_2}, where $1 \leq l_1 \leq l_2 \leq L$. This pairwise comparison could be further transformed into a Boolean function (i.e. $sgn(g_{l_1} - g_{l_2})$)). Together with the L classifiers (one for each cluster), we have a total of $(L+1)L/2$ Boolean functions. Among all these Boolean functions, the maximum VC-dimension is $D+1$, due to $D_{\mathcal{C}} \leq D$. Therefore, by Lemma 2 of Bousquet and Boucheron,[45] the VC-dimension of the function family H is bounded by $2\left(\frac{(L+1)L}{2}\right)(D+1)\log(e\frac{(L+1)L}{2})$ or equivalently $(L+1)L(D+1)\log(e\frac{(L+1)L}{2})$. The generalization guarantees follow from Vapnik.[26]