ABSTRACT

Predicting a target event from temporal data using supervised learning alone presents a number of challenges. It assumes that members falling into the same class have similar historical characteristics, which is a too strong an assumption. Additionally, it can be difficult for the algorithm to underline the differences from a large volume of data and multitude of temporal projections. In such situations, a combination of supervised and unsupervised learning proved to be superior in performance as compared to supervised learning alone. In the proposed methodology, we develop feature vectors of temporal events that are subsequently split into groups by similarity of spatio-temporal characteristics using a clustering algorithm. We then apply a supervised learning methodology to predict the class within each of these subpopulations. We show a dramatic improvement in predictive power of this joint methodology as compared to supervised learning alone. The case study that we use to demonstrate the methodology utilizes medical claims data to predict a patient’s short-term risk of myocardial infarction. In particular, we identify groups of people with temporal diagnostic patterns associated with a high-risk of myocardial infarction in the coming three months. We use these patterns as a profile reference for assessing the state of new patients. We demonstrate that the newly developed combined approach yields improved predictions for myocardial infarction over using classification alone.

INTRODUCTION

Temporal data, which is a collection of observations over time, is fast-growing in availability. Electronic time series data are becoming increasingly common in healthcare as the industry transitions from paper to electronic medical records. Research advances have made DNA sequence transcription and other novel diagnostic methodologies more common. The ability to predict significant events from time series data is an important problem, which includes better diagnostic and patient monitoring abilities, detection of serious complications, and healthcare cost and staffing management among others. Transition to electronic information storage, including medical records, led to the creation of multiple data centers aimed at improving healthcare objectives.

The Naval Health Research Center (NHRC) develops and maintains a medical records database of combat injury and clinical treatment and outcome data for sick and injured personnel during deployment. The database, called the CTR Expeditionary Medical Encounter Database (CTR EMED), was developed to improve readiness for casualty treatment and inform medical providers and planners who allocate supply and personnel needs. This database is an objective source of time series clinical data. The data can be aggregated into temporal events data and then analyzed to reveal patterns of clinical care for US service members sick or injured during deployment (NHRC, 2011).

The Directorate of Personnel Data Systems at the Air Force Personnel Center (AFPC) sustains personnel data and records management systems for the Air Force, Department of Defense (DoD) civilians and more than 100 other federal agencies. Relational databases are under development to support personnel mission requirements at all DoD operational levels. This time series data can be analyzed to discover different personnel groups or characteristics that are more prone to retention or discharge. Predicting retention rates and identifying personnel with future potential for retention or with high risk for discharge can help the DoD downsize to meet National Performance Review goals.

We propose a joint methodology based on a combination of unsupervised clustering with supervised learning for predicting a significant adverse event within temporal data. The methodology improves predictive power over the more traditional method of using supervised learning algorithms alone. Additionally, we develop templates based on multidimensional temporal data that are representative of patterns that are predictive of the event of interest. We demonstrate the effectiveness of the methodology using medical claims data of recent months to predict a patient’s risk of myocardial infarction in the near future. Specifically, by combining clustering and classification algorithms, we built models that use medical claims data.
from more than 47,000 members over five years to calculate the risk of myocardial infarction in real time. In addition, we developed distinctive temporal diagnostic patterns that are associated with higher incidence of myocardial infarction.

The methodology is scalable and can be generalized to multiple sets of problems. It effectively splits the data into groups with different temporal and clinical characteristics, and thus allows us to better uncover dynamic diagnostic patterns in each of the subgroups. In our case study we found that the uncovered diagnostic patterns provide insights into what diagnostic combinations and profiles classify a patient as high-risk for myocardial infarction and are strong predictors of a future myocardial infarction.

The rest of this article is summarized as follows: Next, we describe the analytic approach and background methodologies that form the foundation for our proposed joint methodology. Following that, we present the joint methodology of supervised classification and unsupervised clustering used. In the subsequent, we offer an application in healthcare to demonstrate the joint methodology’s predictive power and ability to uncover multidimensional temporal patterns. Finally, we discuss the significance of the methodology and how it can be generalized for other applications.

RELATED WORK

Because of its breadth of application, determining methods to analyze time series data has been an important emerging area of statistics. Time-series analysis is different from traditional data analysis because the successive observations are dependent, causing the analysis to be based on the time order of the observations. This enables the prediction of future values to be based on historical observations (Chatfield, 2004). Conventional time-series prediction methods, however, are limited and do not consider the prediction of a target event within a window of time. Additionally, many models are built for binary classification, thus requiring the algorithm to uncover different trajectories of data over time leading to the same target event. For example, myocardial infarction can be preceded by a series of cardio-pulmonary events, or by a very different set of anemia and flu-like complications. Not all algorithms are sophisticated enough to incorporate such different temporal patterns, thus leading to weak prediction results. A methodology that successfully uncovers varying combinations of temporal factors associated with or leading to the target event is needed.

Several researchers (Domeniconi et al., 2002; Weiss and Hirsh, 1998; Vilalta et al., 2001) have studied the development of an algorithm to predict target events from time series data. Domeniconi et al. (2002) use singular value decomposition (SVD) combined with support vector machine (SVM) classification to perform co-occurrence analysis of events. Similar to our proposed methodology, the approach starts with an initial rich set of features and then clusters them based on feature correlation. The resulting set is fed to SVM for prediction. However, the work focuses only on predicting a specific event within a window of time, ignoring the member associated with the event. This does not allow for real-time monitoring and prediction for each member. The Timeweaver algorithm by Weiss and Hirsh (1998) is a machine learning system specifically designed to solve rare event prediction problems with categorical features. Similar to our methodology, the algorithm first identifies prediction patterns, and then constructs prediction rules to find the top \( n \) patterns by pruning the bad rules. However, there is little unsupervised nature in the approach, and naturally occurring patterns, uninfluenced by the target event can be missed. Vilalta et al. (2001) uses a methodology similar to Weiss and Hirsh (1998) by generating a set of rules correlating each target event with events occurring previous to the target event within a time window. A prediction model is then run on the pruned training data.

We will improve the developed approaches by providing a better analysis of various temporal characteristics of data that are based on a combination of multiple factors, and that are preceding the target event. We will also develop a real-time approach to assess the risk of the target event occurring in the near future.
PREDICTION IMPROVEMENT

PROBLEM

Approach

There are two different forms of statistical learning: supervised learning and unsupervised learning. In supervised learning, the target event is known and prediction algorithms are used to build a model that predicts the target event from a set of attributes. With unsupervised learning, the target event is unknown and clustering algorithms are used. This form of learning discovers correlations within the set of attributes to find subgroups or patterns inherent in the data. In this work, we use a concurrent combination of prediction and clustering algorithms to predict a Boolean value. Prediction methods alone yield weak results, but a joint approach increases the predictability of the target event significantly. The results are evaluated on the test sets using predictive accuracy measures.

Problem Formulation

The general formulation of the problem is summarized in Figure 1. We assume the data set \( \{x\} \) consists of \( n \) different types of events occurring over an observation period of length \( t \) to \( m \) different members. Event \( x \) is assigned to a member and occurs at a defined time in the observation period. We assume that each member can have multiple events occur in the observation period and/or have the same event occur multiple times in the observation period. The observation period is split into smaller windows of time, \( w_i \) in order to observe events occurring within each window. We compress the occurrence of events corresponding to the same member into one record by mapping the events into vectors, \( e \in \mathbb{R}^a \), where each vector represents a unique member and the dimensionality, \( a \), is given by the number of event types in the data set over each window \( (a = nw) \). Each element, \( e_{w_i} \), is a count of the number of occurrences of the event with respect to the window it occurs in. This creates a time-series representation of the member’s history of events over the observation period which can be used to examine the windows of event history leading up to the occurrence of the target event, \( X \).

The target event, \( X \), is binary and is denoted \( \{ -1, +1 \} \). \( X \) is an event in \( \{x\} \) and can occur anywhere within \( e \). To compare the event histories of all members concurrently, we align the data according to the occurrence of \( X \) to make observations date-independent, thus preserving order of events. The training set, \( T \), is balanced, containing an equal number of positive and negative values for the target event so that neither class receives preferential treatment by the prediction algorithm due to its prevalence in the training set. Negative examples of the target event are randomly selected from the data set from within the same window of time. The test set, \( S \), is generated from the data left over and remains unbalanced so that it represents the original distribution of the occurrence of the target event in the sample population.

JOINT METHODOLOGY

Often, traditional prediction models are used for binary classification. Using these models to predict a target event from temporal data can present a number of challenges. It assumes that members falling into the same prediction class have similar historical characteristics, which is too strong an assumption. Additionally, it can be difficult for the prediction model to underline the differences between classes from a large volume of data and multitude of temporal projections. If the predictive algorithm is not well adjusted to uncover distinctive temporal patterns leading to the target event, it will yield weak predictive results. Sometimes, even highly nonlinear algorithms cannot perform adequately well alone. We will use the Random Forest algorithm, a highly nonlinear and successful classifier, as a baseline method to compare its performance to the proposed joint methodology that successively combines clustering and the random forest classifier. In particular, we first apply clustering algorithms to the data to find different groups of members having certain historical characteristics and patterns. We then predict the target event within each of these groups by developing a prediction model for each cluster. In addition to predicting the target event, we develop templates that describe temporal sequence of events preceding the target
event. The combination of characteristics of members in these groups can help us develop high-risk time-series patterns that are predictive of future target events.

**Clustering Methodology**

Clustering is used in the absence of a target event to search for relationships among a set of features or to organize data points into meaningful groups. It attempts to divide the data into naturally occurring groups, called clusters, so that the data points within each cluster are more similar to each other than to data points in other clusters. We use clustering to reveal correlations among the features associated with the target event. These correlations are used to derive clusters with different temporal patterns and address the notion that there may be many different trajectories of temporal patterns leading to a target event. We use the spectral clustering algorithm in this work and support this choice by comparing it to the more traditional k-means clustering algorithm.

K-means clustering is based on the idea that data can be divided into clusters where each cluster is represented by a center point (Hastie et al., 2009). This center point, also known as the centroid, is the mean of the data within the cluster. To initiate the k-means algorithm, the number of clusters, \( k \), must be specified and the centroid location can be specific or chosen randomly. Given the initial set of \( k \) centroids, the algorithm assigns data points to clusters by minimizing the squared Euclidean distance from the data point to the centroid of the cluster. The centroid, or the mean of the data, is recomputed. The algorithm iteratively re-computes centroids.

**Figure 1.** Summary of the joint methodology.
and re-assigns data points until there is no effective change in cluster assignments.

Many studies have shown that the spectral clustering algorithm (Chen et al., 2010; Luxburg, 2007) is more effective in finding clusters than traditional clustering algorithms such as k-means. Spectral clustering considers the pairwise similarity of data points while k-means only considers the similarity values from the individual data points to their centroids. The algorithm measures the pairwise similarities of the data points and constructs the corresponding similarity matrix. The un-normalized Laplacian is taken and then eigendecomposition is performed. The k-means algorithm is then executed on the resulting matrix. The number of clusters, \( k \), must be specified.

**Prediction Method**

The random forest algorithm is known for its attractive property of detecting variable interactions and excellent performance as a learning algorithm. The algorithm estimates the importance of a variable by looking at how much prediction error increases when data for that variable is permuted while all others are left unchanged (Liaw and Wiener, 2002). When tested on eight different performance metrics against nine other prediction algorithms, random forests gave the second best average performance across all of the metrics and different test problems (Caruana and Niculescu-Mizil, 2006). This establishes our rationale for using a random forest model.

The random forest algorithm builds a large collection of uncorrelated classification trees (Breiman, 2001). A classification tree is a decision tree that recursively partitions a dataset into smaller and smaller groups that are similar based on the known class label. The parameter that requires tuning in this algorithm is \( t \), the number of trees to construct. We built random forest algorithm models using training data and different values of \( t \) to find the optimal value of \( t \). The model with the optimal value of \( t \) was chosen from the results of 10-fold cross-validation: the training set was split into 10 equal parts with nine-tenths of the training set used to train the model and the remaining one-tenth used to calibrate the model and find the optimal value of \( t \).

The independent test set, not used in the model development, is used to evaluate the model’s power to predict the target event on unseen data, thus providing generalization error (Hastie et al., 2009).

**Joint Methodology**

Our method first uses spectral clustering to cluster the training set, \( T \), with the target event values omitted. This creates a model with a set of \( k \) clusters. Each of the \( k \) clusters is a group of members that are described by different patterns of attributes. The data from each of the \( k \) clusters is then used to build \( k \) random forest classification models. These models are used to predict the occurrence of the target event within each cluster separately. After the classification models are trained and validated using 10-fold cross validation, the final model is selected. It is evaluated on the test set, \( S \), that was not used in the model development. \( S \) is run through the cluster model, target event values omitted, and then evaluated by a corresponding classification model. The prediction performance of each classification model is then compared to the baseline performance on the same subset of data to demonstrate the difference in the performance.

**APPLICATION IN HEALTHCARE**

**Scenario**

Myocardial infarction (MI), also known as a heart attack, is a dramatic complication of coronary heart disease that is a growing concern in the United States today. The ambiguity behind some MI events presents the need for understanding patterns in patient’s diagnostic histories that are linked to MI. One way to do this is through the analysis of temporal diagnostic data from administrative databases.

With the data from a patient’s past history of diagnoses we use our joint methodology to uncover combinations of factors associated with or leading to myocardial infarction and to develop models predicting a patient’s risk of myocardial infarction in the near future. Specifically, we apply the joint methodology to medical claims data from more than 47,000 members.
over five years to: 1) develop predictive models for myocardial infarction and evaluate them on out-of-sample groups of patients, 2) find distinctive temporal diagnostic patterns that are associated with higher incidence of myocardial infarction, and 3) show that the joint method is a systematic way for finding patterns already known to lead to MI. We will demonstrate that the proposed methodology yields improved predictive power as compared to using classification alone. The methodology also uncovers multidimensional temporal diagnostic patterns from claims data that are indicative of a higher risk for a patient to develop MI. This is done in a systematic way, providing a global view of disease development.

**Myocardial Infarction Data**

The data is provided by Verisk Health, a company that leverages healthcare data to identify, manage and minimize medical and financial risk at both the patient and population levels. The data is de-identified and is generated from health insurance claims filed for 47,763 members from a commercially insured population across the country over the observation period 01/27/2000–11/30/2007. The data is from members who meet the following criteria:

- Have at least five claims reflecting a coronary artery disease (CAD) diagnosis code and five visits with a hypertension diagnosis code;
- Have complete data for at least a five-year period with continuous eligibility and the most recent data available; and
- Have at least 100 medical claims and at least five pharmacy claims.

These criteria ensure that the study population has a reasonably long and rich history to analyze, with continuous coverage, and that the CAD and hypertension diagnoses are not erroneous. Verisk Health clinicians recommended selecting the CAD/hypertension population to help define a target population with a higher risk of MI so that analysis results would not be sparse. From the health insurance claims, we define an event as a diagnosis or a visit to the hospital. The resulting data set consists of 19,963,685 health insurance entries along with basic demographic information.

**Attribute Aggregation**

The total number of members is 47,763, and we consider 46 different diagnosis groups as event types. The target event is defined as the occurrence of an MI event and a trip to the emergency room with subsequent hospitalization. The observation period is split into 21 windows, each 90 days in length. The size of the window is data dependent and needs to be determined empirically. The proposed length reflected frequency of visits, with the expectation that the majority of period windows were not empty. CAD and hypertensive patients are typically advised to visit their primary care physician every few months in order to track their condition and ensure treatment is occurring as planned. In particular, this window length allows us to better capture abnormal doctor visits and diagnoses outside of the regularly occurring check-ups. A window that is too long will not allow us to effectively capture a trend over time and a window length that is too small will fail to capture meaningful data leading up to the target event. Specifically, the exact point in time that certain diagnoses occur before the target event are dependent on the patient and with small window lengths, this variation can make it difficult to compare diagnostic trajectories of the patients to one another. We count the numbers of medical claims filed for each diagnosis, along with the associated number of visits to the hospital, within each window. We align the data to examine three periods (nine months) of diagnostic history leading up to the target event.

**Results**

Table 1 compares the prediction performance of the joint methodology and the baseline method. The results are reported on the test set not used in the joint model’s development. The first column identifies the cost bucket. In this work, the data is divided into cost buckets to lessen the effects of high-cost outliers and base the analysis on patients with similar conditions of health. See research from Bertsimas et al. (2008) and AMA (2008) for more detail on medical expense analysis and how it can be a good summary of a person’s health. Cost buckets 1 to 3 represent below-average to average health-risk
members, average to above-average health-risk members, and high health-risk members respectively.

The baseline method, which uses only the random forest model, poorly predicts MI. The prediction of MI for members in cost bucket 1 is less than 50 percent and cost bucket 3 is still only 58.3 percent. However, the prediction performance significantly improves with the joint method, where clustering is performed first. Even with \( k = 5 \), the prediction performance of cost bucket 3 reaches 70.39 percent. The joint method’s prediction performance increases with \( k \), performing the best with the \( k = 10 \) cluster model. In this model, cost bucket 2 predicts MI with an accuracy of 71.55 percent and an accuracy of 79.15 percent in cost bucket 3. The joint method improves the prediction performance of MI by more than 30 percent.

### Optimal \( k \) Sensitivity Analysis

In this work we conduct our analysis using a \( k = 10 \) cluster model, which is based on optimal \( k \) sensitivity analysis. The optimal value for \( k \) is usually data dependent and the methodology for finding it is not concrete. The natural number of distinct groupings, \( k^* \), is unknown and must be estimated directly from the data.

One such way for estimating \( k^* \) examines the within-cluster dissimilarity \( W_k \) as a function of the number of clusters \( k \). “Separate solutions are obtained for \( k \in \{1, 2, \ldots, k_{\text{max}}\} \). The corresponding values \( \{W_1, W_2, \ldots, W_{k_{\text{max}}}\} \) generally decrease with increasing \( k^* \) (Hastie et al., 2009). For \( k < k^* \), the cluster model will not adequately divide the members into enough clusters so that data points within each cluster are more similar to each other than to data points in other clusters. Theoretically, \( W_k \) will substantially decrease as \( k \) approaches \( k^* \). For \( k > k^* \), the cluster model will excessively divide the members into clusters, causing one or more of the naturally occurring groups to be partitioned into two subgroups. In this case, \( W_k \) will decrease as \( k \) increases beyond \( k^* \), but the decrease will not be as significant. “That is, \( \{W_k - W_{k+1} | k < k^*\} \gg \{W_k - W_{k+1} | k \geq k^*\}\” (Hastie et al. 2009). We can find an estimate for the optimal \( k^* \) by looking for a kink in the \( W_k \) curve. Figure 2 shows the plot of \( W_k \) as a function of \( k \), with a kink at \( k = 10 \) where there is a smaller decrease in \( W_k \) as \( k \) increases beyond \( k = 10 \).

Another way to estimate \( k^* \) in the joint methodology is to look at the prediction performance \( P_k \) as a function of the number of clusters \( k \). Figure 3 below shows \( P_k \) from the joint method monotonically increasing with \( k \). Theoretically, as \( k \to \infty \), \( P_k \to 1 \). This inherently makes a large \( k \) more desirable. However, the trade-off for increased \( P_k \) is decreased cluster integrity.

In Figure 4, we show the cluster distribution for each of the \( k = 5, 7, 10 \) models in cost bucket 2 to give the reader a visual representation of how the data naturally falls into clusters given different values of \( k \). We observe an uneven distribution of data across the clusters for small values of \( k \) because the majority of the data falls into one large cluster, with the rest of the data divided into the remaining clusters. As we increase to \( k = 7 \), we observe similar results but the distribution of data has improved and the natural sizes of the clusters are more even relative to one another. In the \( k = 10 \) cluster model, there is an improved distribution of data across the clusters without a dominant cluster that contains a majority of the data points. Increasing \( k \) past 10 with the available set of data causes

### Table 1. Prediction performance per cost bucket: Baseline and Joint Method results.

<table>
<thead>
<tr>
<th>Cost bucket</th>
<th>Baseline method (%)</th>
<th>Joint method (with ( k )-clusters) (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>49.63</td>
<td>66.87  67.11  68.25</td>
</tr>
<tr>
<td>2</td>
<td>55.99</td>
<td>60.65  61.43  71.55</td>
</tr>
<tr>
<td>3</td>
<td>58.31</td>
<td>70.39  73.72  79.15</td>
</tr>
</tbody>
</table>

Figure 2. Within-cluster dissimilarity \( W_k \) for different values of \( k \).
the clusters to become very small, thus hampering the predictive power of subsequent supervised learning models that could not be reliably developed on too small of data sets. Too little data in each cluster does not yield enough data to conduct further analysis and therefore we validate our earlier findings that the data does in fact optimally fall into 10 clusters.

Cluster-based Performance Results

Figure 5 shows prediction performance of the $k = 10$ model for each of the cost buckets. We observe an increase in the weighted average prediction performance of MI from cost bucket 1 to cost bucket 3 (68.2 to 79.1 percent) and the individual cluster predictions have less variability around the weighted average. As discussed earlier, cost bucket 3 represents less healthy, more expensive patients. Even though a higher percentage of the members fall into cost buckets 1 and 2, the data in these cost buckets is sparse when compared to data for members in cost bucket 3. This is because members in the first two cost buckets are less sick and therefore have a less dense history of visits to medical facilities. Therefore, less variability around the weighted average prediction performance of MI in cost bucket 3 shows that our model produces better and more stable results with dense data.

It is of interest that the approach found clusters that are strongly indicative of the coming MI event. These clusters not only have high prediction performance but their structures reveal a reasonably large proportion of members with an MI event in the target period. Clusters with a large disparity between the number of patients who experience an MI event in the target period and the number of members who do not are indicative that the pattern of health represented in the cluster is an important predictor of MI. Not pretending to find all such patterns, the methodology is able to automatically capture important trends in claims data that could reliably serve as a warning of coming problems. In particular, the prediction performance for clusters 3, 6, 7, and 9 from cost bucket 3 is over 80 percent, with cluster 8 from this same cost bucket over 90 percent. The predictive accuracy in clusters 2, 7, and 10 from cost bucket 1 exceeds 85 percent and cluster 4 from cost bucket 2 exceeds 82 percent.

Cluster Structures and Patterns

Although patients from different clusters in the same cost bucket may have similar cost characteristics and high probabilities of having an MI event, the diagnostic trajectories are distinct. Many of the health trajectories we found being associated with MI have been previously well known, or have been confirmed by other studies. We found a number of significant temporal patterns preceding the MI event from the clustering models, and from these we selected patterns...
using at least one of the following criteria: large disparity between the number of patients with an MI event and the number without; high prediction performance of an MI event within the cluster; or a significant diagnostic pattern not usually associated with MI. Examples we will expound upon include a pattern of chronic obstructive pulmonary disease (COPD), a pattern of significant chest pain, and a pattern of diabetes. The dynamic diagnostic patterns found in this work can serve as a template to provide insights on what diagnosis combinations and profiles classify a patient as high-risk for MI.

Patterns in Figures 6, 7, and 8 are based on the number of visits to the doctor or to the hospital. Health insurance claims have separate codes and charges for procedures and diagnoses and both of these codes can be present in the same visit to the doctor. Our models are based on diagnosis codes and the only information available is the final diagnosis inserted into the claim after the doctor visit such as the disease, disorder, symptom, and/or medical sign diagnosed. Patients can have multiple visits to the doctor for the same diagnosis in one day because different procedure codes are associated with each visit. In this work, we count each doctor visit separately, even if a patient had several visits per day. This leads to high numbers of visits that could be observed in several patterns. These high numbers might also reflect the concerns doctors have because patients are seriously ill, and therefore data was not modified to exclude procedural visits.

In the text, we refer to events that occurred three, six, and nine months before the onset of MI. As discussed earlier, the data were split into three-month (90-day) periods and each reference corresponds to events from within each period. Specifically, three months before the MI refers to all events that occurred 0–3 months before the time period with the MI event. The six-month period reflects information on events 3–6 months before, and nine months on events 6–9 months before the time period with the MI event. Periods do not overlap. The control population of patients with no MI event follows the same methodology and is selected at random within the same window of time to reflect the same real-time distribution as that found in the population of patients with MI, and to compensate for seasonal and other irregularities. The population of patients with no MI event was much larger than those with an MI event, which enabled selection of the corresponding population.

Figure 6 reflects a pattern in which patients from this cluster, who have an MI target event, regularly visit the doctor for COPD a priori. The solid line shows the average number of visits for each diagnosis for patients with MI, compared against the grey columns, which show the average numbers for these diagnoses in the overall bucket 2. There is a gradual increase in COPD diagnoses with an average of four visits to the doctor nine months before, an average of seven visits six months before, and an average of nine visits within three months before the MI event. The average numbers of visits related to COPD for this cluster are significantly larger than the average number of visits for this diagnosis in the overall bucket 2.
population. The outlined pattern also shows that these patients visit a doctor more frequently than the general population with congestive heart failure (CHF) and anemia diagnoses nine months before the MI, and pneumonia, CHF, and lower respiratory disorders both three and six months before the MI. Although pneumonia is not usually associated with MI, information from Massachusetts General Hospital asserts that warning signs of a heart attack can often be confused with indigestion, pneumonia, pleurisy, or other disorders (Massachusetts General, 2009). An independent study done on the management of COPD patients found that in patients with acute MI, the incidence of COPD was approximately 50 percent higher than in the general population (Behar et al., 1992).

Figure 7 depicts a cluster with a pattern of significant occurrence of chest pain three months before the MI target event. The solid line reflects the average number of visits for each diagnosis for patients from this cluster, compared to the grey columns, which show the average numbers for these diagnoses for the entire population in cost bucket 2. Patients in this cluster have a relatively indistinguishable pattern in the early observational periods. However, three months before the MI, the average number of visits to the doctor with a chest pain diagnosis is increased dramatically from just above 1 to almost 18. Chest pain before a heart attack is a widely accepted diagnosis in medical research and it is satisfactory that the developed automated models successfully uncover known trends.

Figure 8 depicts a pattern of diagnostic characteristics that is well-recognized in the medical community and society. Specifically, it shows a temporal pattern of diabetes and it is a well-known fact that both types 1 and 2 diabetes are associated with accelerated atherosclerosis, one of the main causes of MI (Kulick and Lee, 2010). The solid line reflects the average number

![Figure 6. Pattern of COPD from a cluster in cost bucket 2. Members with an MI target event in this cluster are graphed against the entire cost bucket 2. To enhance the chart’s clarity, labels on the x-axis only include diagnoses with significant values.](image)
of visits for each diagnosis for patients from this cluster, compared to the grey columns, which show the average numbers for these diagnoses for the entire population in cost bucket 3. In this work, we found additional clusters across all cost buckets with other diagnostic patterns that are also well-recognized in the medical community. Well-known diagnoses associated with MI, such as diabetes, hypertension, and hyperlipidemia characterize many of the patterns that consistently occur throughout all of the cost buckets and clustering models. Discovery of these patterns helps to validate our other findings and confirm the strength of our methodology.

CONCLUSION AND FUTURE WORK

Using clustering methods in conjunction with classification algorithms yields improved predictive power over using classification alone. In addition to improved prediction accuracy, clustering methods can effectively split the members into groups with different temporal diagnostic patterns preceding the target event that could serve as visual templates for evaluation of novel input patterns for increased risk of the target event. As found in the example of predicting MI, the patterns found can be a useful profile reference for identifying patients at high risk. Many of the patterns that we found were known from previous independent clinical studies that used different data and methods, thus confirming findings from automated data analysis. We also uncovered novel and previously unseen patterns associated with MI. The proposed approach is scalable and can be generalized and used for various applications that utilize temporal data to provide a systematic way to uncover significant patterns of interest.

The current work was based on utilizing diagnoses for each visit. To further improve the performance and increase the robustness of our findings, future work might include the corresponding procedural and prescription drug information. Different doses and combinations of
drugs over a patient’s clinical history could be indicative of whether the patient’s risk for MI is increasing. This method confirms that claims data can be useful in monitoring patients’ medical conditions and predicting the onset of an adverse event. It can be extended to provide a decision aid real-time support to a doctor by matching patient’s recent records against outlined templates. A match with a pattern associated with MI could serve as a warning for a higher risk of MI for a patient. Because available data are first split into clusters, our approach is limited in that the methodology is better applicable to scenarios with large training data pools. Smaller training sets can be difficult to analyze as the size of clusters will become too small for sufficient supervised learning training. The methodology is also sensitive to the data structure and clustering approach, with less success when a large number of small clusters are created. It is therefore suggested to perform initial data evaluation using different clustering approaches, which will uncover the internal data structure.

The proposed approach can be translated to other diseases and adverse event detection. Current NHRC databases contain a wealth of time series clinical treatment and outcome data from deployed service members. From casualty treatment readiness and supply allocation to detecting high-risk adverse events based on injury or sickness, there exist numerous significant events that if predicted could be an important resource for deployed medical personnel as well as post-deployment treatment. In the combat area of operation, identifying patterns of critical care can help medical personnel provide the right care at the right time to prevent possible death or permanent disability. These patterns can potentially maximize the chances of a soldier returning to duty or civilian life as a healthy and functioning citizen. Post-deployment, patterns of behavior leading to post-traumatic stress disorder (PTSD) and mild traumatic brain injury (mTBI) can be used to evaluate and detect an increased risk for these conditions early on and allow medical intervention. Visual templates derived from temporal patterns can also help sort out the

Figure 8. Medically recognized pattern of diabetes from a cluster in cost bucket 3. Members in this cluster are graphed against the entire cost bucket 3. To enhance the chart’s clarity, labels on the x-axis only include diagnoses with significant values.
complexity of high-risk behavior that has dangerous consequences and identify some of the major factors and combinations of factors that can lead to post-deployment events such as suicide, alcohol-related death, and drug overdose.

AFPC also maintains a database suitable for the proposed approach. With multitudes of temporal data detailing the many attributes of personnel within the Air Force and Department of Defense (DoD), different attributes and combinations of attributes over the duration of a service member’s government employment could be indicative of whether the member is at high risk for discharge. Examples of target events might include being passed over for promotion, developing disqualifying medical issues, Uniform Code of Military Justice (UCMJ) actions for misconduct, or falling short of critical career milestones. This methodology can be extended to provide decision support for current AFPC force reduction efforts by finding patterns that serve as a useful profile reference for identifying personnel with potential for future retention or discharge and are suitable for reduction in force (RIF) measures.

ACKNOWLEDGEMENTS

We would like to thank Sanjay Ghimire and Nathan Gunn at Verisk Health for the extensive data support and Dr. Michael A. Kane for the invaluable clinical expertise and help.

AUTHOR STATEMENT

This article is not a product of PACAF/A9. The views expressed in this article are those of the authors and do not reflect the official policy or position of the Air Force, Department of Defense, or the US Government.

REFERENCES


APPENDIX 1: LIST OF ACRONYMS

AFPC: Air Force Personnel Center
CAD: Coronary artery disease
CHF: Congestive heart failure
COPD: Chronic obstructive pulmonary disease
CTR EMED: CTR Expeditionary Medical Encounter Database
DoD: Department of Defense
MI: Myocardial infarction
NHRC: Naval Health Research Center
SVD: Singular value decomposition
SVM: Support vector machine

APPENDIX 2: BIBLIOGRAPHY