Journal of Machine Learning Research 12 (2011) 999-1024 Submitted 5/10; Revised 11/10; Published 3/11

Unsupervised Similarity-Based Risk Stratification for Cardiovascular Events Using Long-Term Time-Series Data

Zeeshan Syed ZHS@EECS.UMICH.EDU
Department of Electrical Engineering and Computer Science
University of Michigan
Ann Arbor, MI 48109-2121, USA

John Guttag GUTTAG@CSAIL.MIT.EDU
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, MA 02139-4307, USA

Editor: Carla Brodley

Abstract

In medicine, one often bases decisions upon a comparative analysis of patient data. In this paper, we build upon this observation and describe similarity-based algorithms to risk stratify patients for major adverse cardiac events. We evolve the traditional approach of comparing patient data in two ways. First, we propose similarity-based algorithms that compare patients in terms of their long-term physiological monitoring data. Symbolic mismatch identifies functional units in long-term signals and measures changes in the morphology and frequency of these units across patients. Second, we describe similarity-based algorithms that are unsupervised and do not require comparisons to patients with known outcomes for risk stratification. This is achieved by using an anomaly detection framework to identify patients who are unlike other patients in a population and may potentially be at an elevated risk. We demonstrate the potential utility of our approach by showing how symbolic mismatch-based algorithms can be used to classify patients as being at high or low risk of major adverse cardiac events by comparing their long-term electrocardiograms to that of a large population. We describe how symbolic mismatch can be used in three different existing methods: one-class support vector machines, nearest neighbor analysis, and hierarchical clustering. When evaluated on a population of 686 patients with available long-term electrocardiographic data, symbolic mismatch-based comparative approaches were able to identify patients at roughly a two-fold increased risk of major adverse cardiac events in the 90 days following acute coronary syndrome. These results were consistent even after adjusting for other clinical risk variables.

Keywords: risk stratification, cardiovascular disease, time-series comparison, symbolic analysis, anomaly detection

1. Introduction

In medicine, as in many other disciplines, decisions are often based upon a comparative analysis. Patients are given treatments that worked in the past on apparently similar conditions. When given simple data (e.g., demographics, comorbidities, and laboratory values) such comparisons are relatively straightforward. For more complex data, such as continuous long-term signals recorded during physiological monitoring, they are harder. Comparing such time-series is made challenging by three factors: the need to capture the many different changes that occur over long periods, for

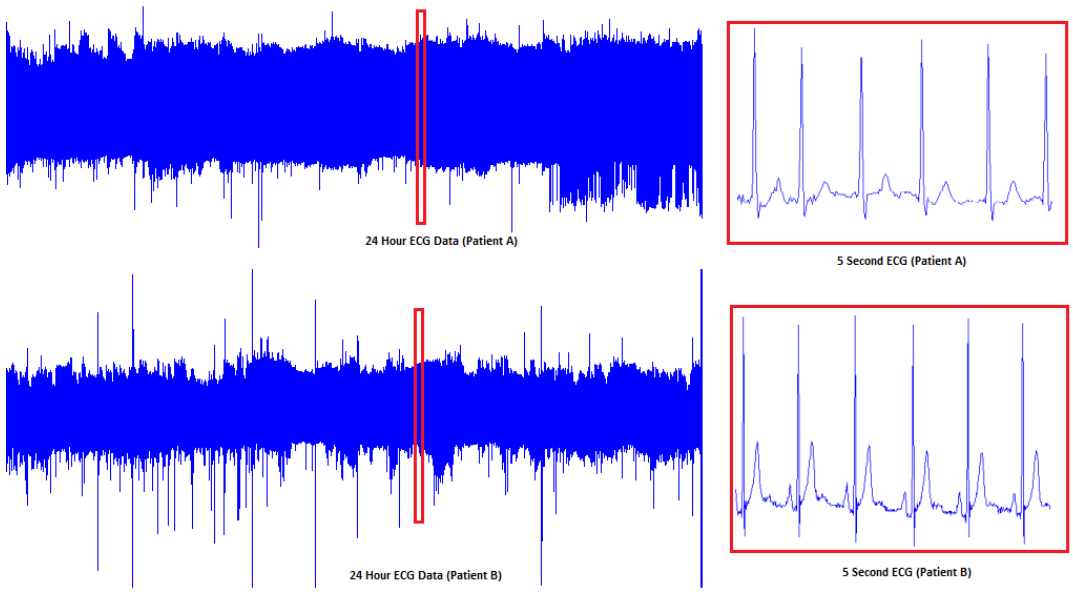

Figure 1: 24 hour ECG signals from two patients. Each time-series is over ten million samples long and contains patient-specific differences in the shape, frequency and time scale of activity over the recording duration. These differences need to be captured while comparing these data.

example, in the shape, frequency, or time scale of activity; the need to efficiently compare very long signals across a large number of patients; and the need to deal with patient-specific differences (Figure 1).

Despite these challenges, comparative analyses of long-term physiological time-series can potentially offer clinically useful prognostic information. While there is an extensive body of research focussed on comparing relatively short time-series, including measures such as dynamic time warping (Keogh and Pazzani, 2001; Keogh and Ratanamahatana, 2005), longest common subsequence (Vlachos et al., 2002), edit distance with real penalty (Cheng and Ng, 2004), sequence weighted alignment (Morse and Patel, 2007), spatial assembling distance (Chen et al., 2007), this work does not directly focus on comparing very long time-series (e.g., millions of samples long). In this paper, we investigate the hypothesis that comparative analyses of long-term physiological activity can aid risk stratification and present symbolic mismatch as a way to quantify differences between the physiological signals of patients. Symbolic mismatch compares long-term time-series by mapping the original signals into a symbolic domain and quantifying differences between the morphology and frequency of prototypical functional units. The use of symbolization is an abstraction process that greatly reduces the size of the data to be compared. For example, comparing the long-term electrocardiographic (ECG) activity between two patients may involve comparing over a hundred thousand beats (with each beats having roughly 500 samples per beat). Using symbolization to reduce this data to a small number of representative units can greatly decrease the size of this problem. This reduction allows for symbolic mismatch to be useful in analyzing very long time-series.

We describe how this measure can be modified in a reasonably simple manner for use with kernel classifiers.

We also present different ways in which symbolic mismatch can be used to identify high risk patients in a population. The methods we propose are motivated by the observation that high risk patients typically constitute a small minority in a population. For example, cardiac mortality over a 90 day period following acute coronary syndrome (ACS) was reported to be 1.79% for the SYMPHONY trial involving 14,970 patients (Newby et al., 2003) and 1.71% for the DISPERSE2 trial with 990 patients (Cannon et al., 2007). The rate of myocardial infarction (MI) over the same period for the two trials was 5.11% for the SYMPHONY trial and 3.54% for the DISPERSE2 trial. Our hypothesis is that these patients can be discovered as anomalies in the population, that is, their physiological activity over long periods of time is dissimilar to the majority of the patients in the population. In contrast to algorithms that require labeled training data, we propose identifying these patients using unsupervised approaches based on three methods previously reported in the literature: one-class support vector machines (SVMs) (Scholkopf et al., 2001), nearest neighbor analysis (Cover and Hart, 1967) and hierarchical clustering (Ward Jr, 1963).

In the remainder of this paper, we describe our work in the context of risk stratification for cardiovascular disease. Cardiovascular disease is the leading cause of death globally and causes roughly 17 million deaths each year (World Health Organization, 2009). By 2030, this number is expected to grow to nearly 24 million deaths annually (World Health Organization, 2009). Most of these cases are expected to be the result of coronary attacks. Despite improvements in survival rates, in the United States, 1 in 4 men and 1 in 3 women still die within a year of a recognized first heart attack (Mackay et al., 2004). This risk of death can be substantially lowered with an appropriate choice of treatment (e.g., drugs to lower cholesterol and blood pressure, aspirin; operations such as coronary artery bypass graft and balloon angioplasty; and medical devices such as pacemakers and implantable cardioverter defibrillators) (World Health Organization, 2009). However, matching patients with treatments that are appropriate for their risk has proven to be challenging (Bailey et al., 2001; Lopera and Curtis, 2009). Existing techniques based upon conventional medical knowledge continue to be inadequate for risk stratification. This leads us to explore methods with few *a priori* assumptions. We focus, in particular, on identifying patients at risk of major adverse cardiac events (death, myocardial infarction and severe recurrent ischemia) following coronary attacks. This work uses long-term ECG signals recorded during patient admission for acute coronary syndrome. These signals are routinely collected, potentially allowing for the work presented here to be deployed easily without imposing additional needs on patients, caregivers, or the healthcare infrastructure. We demonstrate that it is possible to do long-term ECG-based risk stratification without defining a set of features, performing cross-patient symbol or feature alignment, or having any labeled data.

The main contributions of our work are: (1) we describe a novel approach for cardiovascular risk stratification that is complementary to existing clinical approaches, (2) we explore the idea of similarity-based clinical risk stratification where patients are categorized in terms of their similarities rather than specific features based on prior knowledge, (3) we develop the hypothesis that patients at future risk of adverse outcomes can be detected using an unsupervised approach as outliers in a population, (4) we present symbolic mismatch, as a way to efficiently compare very long time-series without first reducing them to a set of features or requiring symbol registration across patients, and (5) we present a rigorous evaluation of unsupervised similarity-based risk stratifying using long-term data in a population of 700 patients with detailed admissions and follow-up data.

While this manuscript focuses on cardiovascular disease, we believe that the ideas presented here can potentially be extended to other data sets in a relatively straightforward manner.

2. Background

We start by briefly reviewing the clinical background for our research. We focus, in particular, on aspects of cardiac function related to electrophysiology. This is followed by a discussion of acute coronary syndromes (ACS) and a summary of recently proposed long-term ECG metrics for risk stratification following ACS. Finally, we present a short overview of survival analysis methods used to evaluate metrics for risk stratification.

2.1 Electrocardiogram

An electrocardiogram (ECG) is a continuous recording of the electrical activity of the heart muscle or myocardium (Lilly, 2007). A cardiac muscle cell at rest maintains a negative voltage with respect to the outside of the cell. During depolarization, this voltage increases and may even become positive. Consequently, when depolarization is propagating through a cell, there exists a potential difference on the membrane between the part of the cell that has been depolarized and the part of the cell at resting potential. After the cell is completely depolarized, its membrane is uniformly charged again, but at a more positive voltage than initially. The reverse situation takes place during repolarization, which returns the cell to baseline.

These changes in potential, summed over many cells, can be measured by electrodes placed on the surface of the body. For any pair of electrodes, a voltage is recorded whenever the direction of depolarization (or repolarization) is aligned with the line connecting the two electrodes. The sign of the voltage indicates the direction of depolarization, and the axis of the electrode pair is termed the lead. Multiple electrodes along different axes can be used so that the average direction of depolarization, as a three-dimensional vector, can be reconstructed from the ECG tracings. However, such multi-lead data is not always available, especially for portable ECG monitors that maximize battery life by reducing the number of electrodes used. Much of our work is therefore designed for the single ECG lead case. As we show subsequently, there is sufficient information even within a single lead of ECG to risk stratify patients.

ECG data is routinely recorded for hospitalized patients, since it is useful for determination of heart rate and pulse, and for the detection of progressive cardiac disease and complicating arrhythmias (Goldstein, 1997). ECG has the advantage of being easy to acquire; the electrical activity of the heart can be measured on the surface of the body in an inexpensive and non-invasive manner. In an in-patient setting, the ECG is typically captured by bedside monitors. In an out-patient setting, a Holter monitor (a portable ECG device worn by patients) can record data continuously over multiple days.

The ECG is a quasi-periodic signal (i.e., corresponding to the quasi-periodic nature of cardiac activity). Three major segments can be identified in a normal ECG. The *P wave* is associated with depolarization of cardiac cells in the upper two chambers of the heart (i.e., the atria). The *QRS complex* (comprising the Q, R and S waves) is associated with depolarization of cardiac cells in the lower two chambers of the heart (i.e., the ventricles). The *T wave* is associated with repolarization of the cardiac cells in the ventricles. The QRS complex is larger than the P wave because the ventricles are much larger than the atria. The QRS complex also coincides with the repolarization of the atria, which is therefore usually not seen on the ECG. The T wave has a larger width and

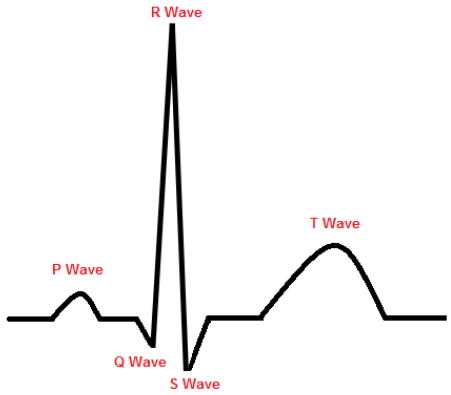

Figure 2: Schematic representation of the normal ECG for a single heart beat.

smaller amplitude than the QRS complex because repolarization takes longer than depolarization (Lilly, 2007). Figure 2 presents a schematic representation of the normal ECG.

2.2 Acute Coronary Syndromes

In our research, we mainly focus on identifying high risk patients following acute coronary syndrome (ACS), an umbrella term covering clinical symptoms compatible with reduced blood supply to the heart (i.e., myocardial ischemia). Heart attacks and unstable angina are included in this group. An ACS is an event in which the blood supply to the myocardium is blocked or severely reduced. The most common symptom of ACS is unusual and unprovoked chest pain, but this may often be absent (most notably in patients with diabetes who experience "silent" heart attacks). Other symptoms may include shortness of breath, profuse sweating, and nausea.

An ACS is usually caused by the rupture of an atherosclerotic plaque producing a clot within a coronary artery. This restricts blood flow to the heart, causing ischemia and potentially cell death in the myocardium. Various sub-classifications of ACS are distinguished by the presence of myocardial necrosis (cell death) and by ECG diagnosis (Lilly, 2007).

Unstable angina refers to an ACS event in which necrosis does not occur, while *myocardial infarction* (MI) refers to one in which it does. Usually, we distinguish between ACS where the ECG shows an ST segment elevation (indicative of complete occlusion of an artery with myocardial necrosis) and ACS where the ECG does not show an ST segment elevation (indicative of partal occlusion of an artery with or without myocardial necrosis). Patients with ST segment elevation are given the diagnosis of ST-elevation MI (STEMI) while patients without ST segment elevation are given the diagnosis of non-ST-elevation ACS (NSTEACS). NSTEACS can correspond to either unstable angina or non-ST-elevation MI (NSTEMI). Patients with STEMI are typically at a higher risk relative to patients with NSTEACS. The differentiation between whether the NSTEACS corresponds to unstable angina or NSTEMI is done on the basis of whether necrosis occurs. This involves blood tests to measure the levels of two serum biomarkers, cardiac-specific troponin and creatine kinase MB, which are chemicals released into the bloodstream when myocardial cells die.

2.3 Post-ACS Risk Stratification and Long-Term ECG Techniques

Since patients who experience ACS remain at an elevated risk of death even after receiving treatment for the initial index event (Newby et al., 2003), post-ACS risk stratification is an important clinical step in determining which patients should be monitored and treated more (or less) aggressively subsequently. Roughly speaking, the goal of risk stratification is to identify groups of patients within the post-ACS population with different rates of adverse outcomes despite receiving similar care. This information can provide an important basis to deliver customized care and to perform clinical cost-benefit analyses.

A number of different methods have been proposed to risk stratify patients following ACS with an excellent review provided by Breall and Simons (2010) and Alpert (2010). We focus here on recent techniques for risk stratification based on information in long-term ECG. A variety of methods have been proposed that assess risk based on automated analysis of long-term ECG data collected in the hours or days following admission. Such data is routinely collected during a patient's stay and therefore these risk assessments can be obtained at almost no additional cost. We discuss three ECG-based methods that have been proposed in the literature: *heart rate variability* (HRV) (Malik et al., 1996; Kleiger et al., 2005), *heart rate turbulence* (HRT) (Barthel et al., 2003), and *deceleration capacity* (DC) (Bauer et al., 2006). Each of these measures has been shown to correlate with an increased risk of adverse events in the period following an ACS. One additional long-term ECG-based risk stratification technique, T-wave alternans (TWA) (Smith et al., 1988; Rosenbaum et al., 1994), has also shown promise. However, evaluating TWA typically requires the use of specialized recording equipment, patient maneuvers or chemical agents to elevate heart rate, and input by a human expert for signal selection. As a result we do not consider TWA in our present study, while focusing on automated methods that can be applied to data collected routinely from post-ACS patients.

The class of ECG-based risk stratification techniques that has been discussed most extensively in the literature is HRV (Malik et al., 1996; Kleiger et al., 2005). The theory underlying HRV-based techniques is that in healthy people, the body should continuously compensate for changes in oxygen demand by changing the heart rate. The heart rate should also change as a result of physiological phenomena such as respiratory sinus arrhythmia (Lilly, 2007). A heart rate that changes little suggests that the heart or its control systems are not actively responding to stimuli.

HRT (Barthel et al., 2003) is related to HRV in that it studies the autonomic tone of patients (i.e., control of the heart rate by the nervous system). HRT studies the return to equilibrium of the heart rate after premature beats. Typically, following a premature beat there is a brief speed-up in heart rate followed by a slow decrease back to the baseline rate. This corresponds to the "turbulence" in the heart rate and is present in patients with a healthy autonomic nervous system. HRT is essentially a reflex phenomenon. When a premature beat interrupts the normal cardiac cycle, the ventricles have not had time to fill to their normal levels, resulting in a weaker pulse. This triggers pressure reflex mechanisms that compensate by increasing the heart rate. This compensatory increase in heart rate causes blood pressure to overcompensate and actives the pressure reflex in reverse. Patients that have diminished HRT responses after premature beats are believed to be at high risk due to abnormal nervous control of the cardiovascular system.

DC (Bauer et al., 2006) is an extension of work on HRV and HRT, and also studies information in the heart rate signal. The theory underlying DC is that decreased decelerations in the heart rate suggest an increased unresponsiveness of the heart to cardioprotective signals from the vagal system

for the heart to slow down. This is often the first line of defense against major adverse events such as fatal arrhythmias.

2.4 Survival Analysis

Metrics for risk stratification are typically evaluated using survival analysis techniques. These methods study the rates of adverse outcomes in patients assigned to different groups (e.g., high and low risk groups in the context of risk stratification post-ACS). In general, such analyses can be carried out in terms of sensitivity and specificity. However, data from clinical studies often consists of a mix of patients who either remain event free throughout the duration of the study, experience events at fairly different times within the study, or leave the study before it is complete (a phenomenon termed censoring). Survival analysis focuses on using information in all these cases, that is, by factoring in both the timing of events in patients who experience adverse outcomes, and by using data from patients who leave the study early for the period during which they participated.

Survival data is commonly visualized as a Kaplan-Meier survival curve (Efron, 1988), which reflects the event rate over time in patients belonging to different groups. We present examples of Kaplan-Meier survival curves subsequently in this manuscript. Visually observed differences between Kaplan-Meier survival curves (i.e., differences in the rates of events over time in patients belonging to two or more groups) can be quantified through a variety of methods. The Cox proportional hazards test is most widely used (Cox, 1972) and corresponds to a regression-based approach to determine how the relative risk between populations varies over time in response to explanatory covariates. The outcomes of this analysis are typically a hazard ratio estimating the multiplicative increase in the rate of events over time between populations, and a measure of the statistical strength of this estimate (a 95% confidence interval for the hazard ratio or a p value). Findings are usually considered to be significant if the p value is less than 0.05 (corresponding to a 5% chance of rejecting the null hypothesis, that is, a Type I error).

3. Symbolic Mismatch

In this section, we describe the process through which symbolic mismatch is computed between a pair of long-term ECG time-series. The use of symbolic mismatch for risk stratification is presented in Section 4.

3.1 Symbolization

As a first step, the ECG signal z_m for each patient $m = 1, ..., n$ is symbolized using the technique proposed by Syed et al. (2007). To segment the ECG signal into beats, we use two open-source QRS detection algorithms (Hamilton and Tompkins, 1986; Zong et al., 2003). QRS complexes are marked at locations where both algorithms agree. A variant of dynamic time-warping (DTW) (Myers and Rabiner, 1981) is then used to quantitatively measure differences in morphology between beats. Using this information, beats with distinct morphologies are partitioned into groups, with each group being assigned a unique label or symbol. This is done by means of a Max-Min iterative clustering algorithm that starts by choosing the first observation as the first centroid, c_1, and initializes the set S of centroids to $\{c_1\}$. During the i-th iteration, c_i is chosen such that it maximizes the minimum difference between c_i and observations in S:

$$c_i = \arg\max_{x \notin S} \min_{y \in S} C(x,y)$$

where $C(x,y)$ is the DTW difference between x and y. The set S is incremented at the end of each iteration such that $S = S \cup c_i$.

The number of clusters discovered by Max-Min clustering is chosen by iterating until the maximized minimum difference falls below a threshold θ (chosen empirically as the 'knee' of the maximized minimum difference curve). At this point, the set S comprises the centroids for the clustering process, and the final assignment of beats to clusters proceeds by matching each beat to its nearest centroid. Each set of beats assigned to a centroid constitutes a unique cluster. The final number of clusters, γ, obtained using this process depends on the separability of the underlying data.

Intuitively, the Max-Min clustering algorithms attempts to partition the data into groups such that the similarity of points within the same groups is minimized, while the similarity of points within different groups is maximized. Theoretical analyses of Max-Min cut-based methods show that this approach leads to more balanced separations of the data than other approaches (Ding et al., 2001).

The overall effect of DTW-based partitioning of beats is to transform the original raw ECG signal into a sequence of symbols, that is, into a sequence of labels corresponding to the different beat morphology classes that occur in the signal. Our approach differs from the methods typically used to annotate ECG signals in two ways. First, we avoid using specialized knowledge to partition beats into known clinical classes. There is a set of generally accepted labels that cardiologists use to differentiate distinct kinds of heart beats. For example, the Association for the Advancement of Medical Instrumentation (AAMI) standards define five basic beat classes (AAMI, 1998, 1987; de Chazal et al., 2004), while the Physionet MIT-BIH Arrhythmia database sub-divides these classes to produce 18 different heart beat labels (Physionet, 2005). However, in some cases, even finer distinctions than provided by these labels can be clinically relevant (Syed et al., 2007). Our use of beat clustering rather than beat classification allows us to identify a set of characteristic morphology classes dynamically from the data that capture these finer-grained distinctions. Second, our approach does not involve extracting features (e.g., the length of the beat or the amplitude of the P wave) from individual beats. Instead, our clustering algorithm compares the entire raw morphology of pairs of beats. This approach is potentially advantageous, because it does not assume *a priori* knowledge about what aspects of a beat are most relevant. It can also be easily extended to other time-series data (e.g., blood pressure waveforms and respiratory activity).

3.2 Measuring Mismatch in Symbolic Representations

Denoting the set of symbol centroids for patient p as S_p and the set of frequencies with which these symbols occur in the electrocardiogram as F_p (for patient q an analogous representation is adopted), we define the symbolic mismatch between the long-term ECG time-series for patients p and q as:

$$\Psi_{p,q} = \sum_{p_i \in S_p} \sum_{q_j \in S_q} C(p_i, q_j) F_p[p_i] F_q[q_j] \tag{1}$$

where $C(p_i, q_j)$ corresponds to the DTW cost of aligning the centroids of symbol classes p_i and q_j.

Intuitively, the symbolic mismatch between patients p and q corresponds to an estimate of the expected difference in morphology between any two randomly chosen beats from these patients.

The symbolic mismatch computation achieves this by weighting the difference between the centroids for every pair of symbols for the patients by the frequencies with which these symbols occur.

An important feature of symbolic mismatch is that it is explicitly designed to avoid the need to set up a correspondence between the symbols of patients p and q. In contrast to cluster matching techniques (Chang et al., 1997; Cohen and Richman, 2002) to compare data for two patients by first making an assignment from symbols in one patient to the other, symbolic mismatch does not require any cross-patient registration of symbols. Instead, it performs weighted morphologic comparisons between all symbol centroids for patients p and q. As a result, the symbolization process does not need to be restricted to well-defined labels and is able to use a richer set of patient-specific symbols that capture fine-grained activity over long periods of time.

3.3 Spectrum Clipping

The formulation for symbolic mismatch in Equation 1 gives rise to a symmetric dissimilarity matrix. For methods that are unable to work directly from dissimilarities, this can be transformed into a similarity matrix using a generalized radial basis function (Vanschoenwinkel and Manderick, 2005). For both the dissimilarity and similarity case, however, symbolic mismatch can produce a matrix that is indefinite. This can be problematic when using symbolic mismatch with kernel-based algorithms because the optimization problems become non-convex and the underlying theory is invalidated (Chen et al., 2009b). In particular, kernel-based classification methods require Mercer's condition (Scholkopf and Smola, 2002) to be satisfied by a positive semi-definite kernel matrix. This creates the need to transform the symbolic mismatch matrix before it can be used as a kernel in these methods.

We use spectrum clipping to generalize the use of symbolic mismatch for classification. This approach has been shown both theoretically and empirically to offer advantages over other strategies (e.g., spectrum flipping, spectrum shifting, spectrum squaring, and the use of indefinite kernels) (Chen et al., 2009a). The symmetric mismatch matrix Ψ has an eigenvalue decomposition:

$$\Psi = U^T \Lambda U$$

where U is an orthogonal matrix and Λ is a diagonal matrix of real eigenvalues:

$$\Lambda = diag(\lambda_1, ..., \lambda_n).$$

Spectrum clipping makes Ψ positive semi-definite by clipping all negative eigenvalues to zero. The modified positive semi-definite symbolic mismatch matrix is then given by:

$$\Psi_{clip} = U^T \Lambda_{clip} U$$

where:

$$\Lambda_{clip} = diag(max(\lambda_1, 0), ..., max(\lambda_n, 0)).$$

Using Ψ_{clip} as a kernel matrix is then equivalent to using $(\Lambda_{clip})^{1/2} u_i$ as the representation of the i-th training sample.

We note that while we introduce spectrum clipping mainly for the purpose of broadening the applicability of symbolic mismatch to kernel-based methods, this approach offers additional advantages. Some researchers, for example, assume that negative eigenvalues of the similarity matrix are

caused by noise and view spectrum clipping as a denoising step (Wu et al., 2005). The results of our experiments, presented later in this paper, support the view of spectrum clipping being useful in a broader context.

4. Risk Stratification Using Symbolic Mismatch

We now present three different approaches using symbolic mismatch to identify high risk patients in a population. We choose these algorithms consistent with Eskin et al. (2002), as methods that can operate on high dimensional data, and that each detect points lying in sparse regions of the feature space in different ways. The first of these approaches uses a one-class SVM and a symbolic mismatch similarity matrix obtained using a generalized radial basis transformation. The other two approaches, nearest neighbor analysis and hierarchical clustering, use the symbolic mismatch dissimilarity matrix. In each case, the symbolic mismatch matrix is processed using spectrum clipping. All three of our approaches focus on finding patients with long-term ECG time-series that are anomalies in the population.

4.1 Classification Approach

Our first approach is based on SVMs. SVMs have been applied to anomaly detection in the one-class setting. Scholkopf et al. (2001) propose a method of adapting the SVM methodology to the one-class classification problem. This is done by mapping the data into the feature space corresponding to the kernel and separating instances from the origin with the maximum margin. To separate data from the origin, the following quadratic program is solved:

$$\min_{w,\xi,p} \frac{1}{2}\|w\|^2 + \frac{1}{vn}\sum_i \xi_i - p$$

subject to:

$$(w \cdot \Phi(z_i)) \geq p - \xi_i \;\; i = 1,...,n \;\; \xi_i \geq 0$$

where v reflects the tradeoff between incorporating outliers and minimizing the size of the support region.

For a new instance, the label is determined by evaluating which side of the hyperplane the instance falls on in the feature space. The resulting predicted label in terms of the Lagrange multipliers α_i and the spectrum clipped symbolic mismatch similarity matrix Ψ_{clip} is then:

$$\hat{y}_j = sgn(\sum_i \alpha_i \Psi_{clip}(i,j) - p).$$

We apply this approach to train a one-class SVM on all patients. Patients who lie outside the enclosing boundary are then labeled as anomalies. The parameter v can be varied to control the size of this group.

4.2 Nearest Neighbor Approach

Our second approach is based on the concept of nearest neighbor analysis. The assumption underlying this approach is that normal data instances occur in dense neighborhoods, while anomalies occur far from their closest neighbors.

We use an approach similar to that in Eskin et al. (2002). In this case, the anomaly score of each patient's long-term time-series is computed as the sum of its distances from time-series for its *k*-nearest neighbors, as measured by symbolic mismatch. Patients with anomaly scores exceeding a threshold θ are labeled as anomalies.

4.3 Clustering Approach

Our third approach is based on agglomerative hierarchical clustering. We begin by putting each patient in a separate cluster, and then proceed in each iteration to merge the two clusters that are most similar to each other. The distance between two clusters is defined as the average of the pairwise symbolic mismatch of the patients in each cluster. The clustering process terminates when it enters a region of diminishing returns (i.e., at a 'knee' of the curve corresponding to the distance of clusters merged together at each iteration). At this point, all patients outside the largest cluster are labeled as anomalies.

5. Evaluation Methodology

We evaluated our work on patients enrolled in the DISPERSE2 trial (Cannon et al., 2007). Patients in the study were admitted to a hospital with non-ST-elevation acute coronary syndrome (non-ST-elevation myocardial infarction or unstable angina). Three lead continuous ECG monitoring (LifeCard CF / Pathfinder, DelMar Reynolds / Spacelabs, Issaqua WA) was performed for a median duration of 4 days at a sampling rate of 128 Hz. The endpoints of cardiovascular death, myocardial infarction and severe recurrent ischemia were adjudicated by a blinded panel of clinical experts for a median follow-up period of 60 days. The maximum follow-up was 90 days. Data from 686 patients was available after removal of noise-corrupted signals. During the follow-up period there were 14 cardiovascular deaths, 28 myocardial infarctions, and 13 cases of severe recurrent ischemia. We define a major adverse cardiac event to be any of these three adverse events. The clinical characteristics of this patient population are presented in Table 1.

We studied the ability of each approach (i.e., classification, nearest neighbor analysis and clustering) to identify a high risk group of patients. Consistent with earlier studies to evaluate new methods for risk stratification in the setting of acute coronary syndrome (Shlipak et al., 2008), we classified patients in the highest quartile as the high risk group. For the classification approach, this corresponded to choosing ν such that the group of patients lying outside the enclosing boundary constituted roughly 25% of the population. For the nearest neighbor approach we investigated all odd values of k from 3 to 9, and patients with anomaly scores in the top 25% of the population were classified as being at high risk. For the clustering approach, the varying sizes of the clusters merged together at each step made it difficult to select a high risk quartile. Instead, patients lying outside the largest cluster were categorized as being at risk. In the tests reported later in this paper, this group contained roughly 23% the patients in the population. We used the LIBSVM implementation for our one-class SVM. Both the nearest neighbor and clustering approaches were carried out using MATLAB implementations.

We employed Kaplan-Meier survival analysis (Efron, 1988) to compare the rates for major adverse cardiac events between patients declared to be at high and low risk for all three approaches. Hazard ratios (HR) and 95% confidence interval (CI) were estimated using a Cox proportional hazards regression model (Cox, 1972). The predictions of each approach were studied in univariate models, and also in multivariate models that additionally included other clinical risk variables

Age (years)	62 (53 to 70)
Age≥65 years	41%
Female Gender	36%
Current Smoker	57%
Hypertension	69%
Diabetes Mellitus	23%
Hyperlipidemia	64%
History of COPD	9%
History of CHD	37%
Previous MI	25%
Previous angina	58%
ST depression>0.5mm	66%
Index diagnosis of MI	49%

Table 1: Clinical characteristics of patient population used for study.

(age≥65 years, gender, smoking history, hypertension, diabetes mellitus, hyperlipidemia, history of chronic obstructive pulmonary disorder (COPD), history of coronary heart disease (CHD), previous MI, previous angina, ST depression on admission, index diagnosis of MI) as well as ECG risk metrics proposed in the past such as heart rate variability (HRV) (Malik et al., 1996), heart rate turbulence (HRT) (Barthel et al., 2003), and deceleration capacity (DC) (Bauer et al., 2006).

For HRV, we used metrics proposed by the Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology: the standard deviation of normal-to-normal intervals (SDNN), standard deviation of sequential five minute normal-to-normal means (SDANN), mean of the standard deviation of sequential five minute normal-to-normal intervals (ASDNN), root mean square successive differences (rMSSD), heart rate variability index (HRVI), percent of normal-to-normal interval increments greater than 50 ms (pNN50) and the ratio of low frequency power to the high frequency power (LF/HF). While we computed all HRV measures, we only report results for the best performing one, that is, LF/HF. HRV-LF/HF was dichotomized at 0.95 using the results reported earlier in the literature (Malik et al., 1996).

We measured HRT and DC for each patient using the libRASCH software shared for research use by the inventors of the method (Technische Universitat Munchen, Munich, Germany). HRT was trichotomized based on the turbulence onset (TO) and turbulence slope (TS) as follows: HRT0 (TO<0 and TS>2.5ms), HRT1 (either TO>0 or TS<2.5ms), and HRT2 (TO>0 and TS<2.5ms) (Barthel et al., 2003). DC was trichotomized as follows: category 0 (>4.5ms), category 1 (2.5 ms-4.5 ms), and category 2 (<2.5ms) (Bauer et al., 2006). Both HRT and DC were treated as continuous variables in our study.

We did not compare our work to T-wave alternans (TWA) (Rosenbaum et al., 1994) as TWA is typically measured using specialized hardware and maneuvers to increase the heart rate. While we experimented with a TWA algorithm that has recently been proposed to measure TWA on ECG data collected routinely during admission (Nearing and Verrier, 2002), this algorithm did not produce good results. We believe these results reflect an inability to measure TWA without specialized hardware and manoeuvres, as opposed to the lack of predictive discrimination for the method. We therefore excluded TWA from our analysis, as an ECG approach that is more appropriate for ECG signals collected under specialized conditions.

Method	HR	P Value	95% CI
One-Class SVM	1.38	0.033	1.04-1.89
3-Nearest Neighbor	1.91	0.031	1.06-3.44
5-Nearest Neighbor	2.10	0.013	1.17-3.76
7-Nearest Neighbor	2.28	0.005	1.28-4.07
9-Nearest Neighbor	2.07	0.015	1.15-3.71
Hierarchical Clustering	2.04	0.017	1.13-3.68

Table 2: Univariate association of high risk predictions from different approaches using symbolic mismatch with major adverse cardiac events over a 90 day period following acute coronary syndrome.

6. Results

We divide our results into two broad groups: univariate results (symbolic mismatch-based approaches in univariate models), and multivariate results (symbolic mismatch-based approaches in multivariate models). We also report on the effect of spectrum clipping on performance and provide some brief results regarding the runtime performance of our approach.

6.1 Univariate Results

Results of univariate analysis for all three unsupervised symbolic mismatch-based approaches are presented in Table 2. The predictions from all methods showed a statistically significant (i.e., $p < 0.05$) association with major adverse cardiac events following acute coronary syndrome. The results in Table 2 can be interpreted as roughly a doubled rate of adverse outcomes per unit time in patients identified as being at high risk by the nearest neighbor and clustering approaches. For the classification approach, patients identified as being at high risk had a nearly 40% increased risk of adverse outcomes. Kaplan-Meier survival curves for all three methods are shown in Figures 3 to 5. For the nearest neighbor approach, we present only the results for the best performing k (i.e., $k = 7$).

For comparison, we also include the univariate association of the other clinical and ECG risk variables in our study (Tables 3 and 4). Both the nearest neighbor and clustering approaches had a higher hazard ratio in this patient population than any of the clinical and ECG risk variables studied. Of the clinical risk variables, only age was found to be associated on univariate analysis with major cardiac events after acute coronary syndrome. Diabetes (p=0.072) was marginally outside the 5% level of significance. Of the ECG risk variables, both HRT and DC showed a univariate association with major adverse cardiac events during follow-up.

These results suggest that unsupervised risk stratification using symbolic mismatch can successfully identify patients at an elevated risk of major adverse cardiac events following ACS. In particular, our data shows that patients categorized as high risk by our methods continue to experience an increased risk of adverse outcomes throughout the entire 90 day period post-ACS (Figures 3-5). Our findings are also encouraging in that the relative increase in patient risk found using our methods compares quite favorably with other metrics based on specialized knowledge that are used in existing cardiac risk scoring tools. While the size of our patient population leads us to avoid statements about the nearest neighbor and clustering approaches being better than the other variables in our study (i.e., on the basis of having a higher observed hazard ratio than these other variables), we

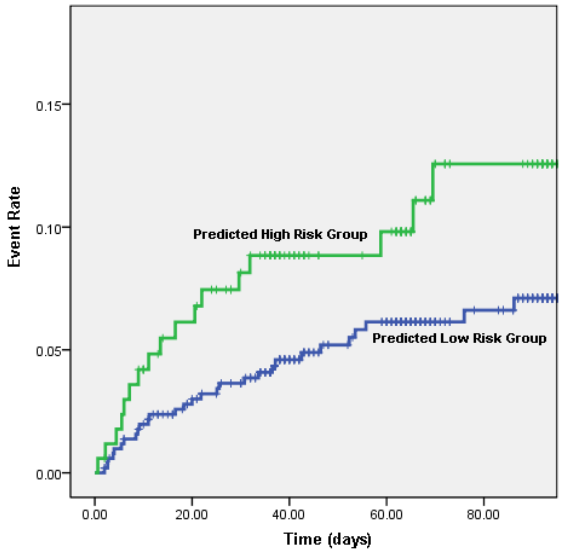

Figure 3: Kaplan-Meier major cardiac event curve for the one-class SVM approach. Ticks represent patient censoring (i.e., patients leaving the study before completion). The top (green) line corresponds to patients with anomaly scores in the top quartile of the population.

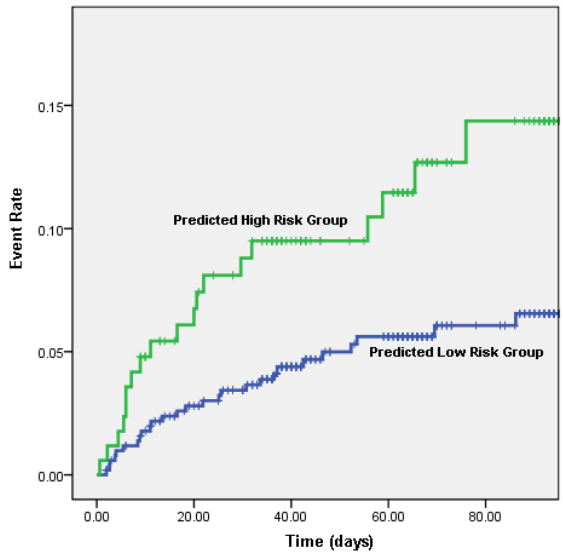

Figure 4: Kaplan-Meier major cardiac event curve for the 7-nearest neighbor approach. Ticks represent patient censoring (patients leaving the study before completion). The top (green) line corresponds to patients with anomaly scores in the top quartile of the population.

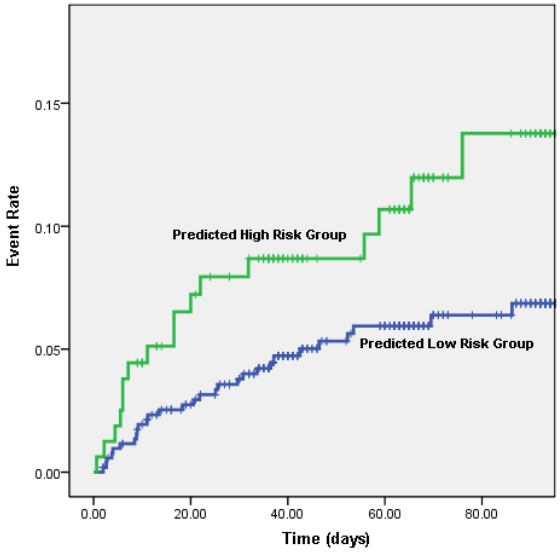

Figure 5: Kaplan-Meier major cardiac event curve for the hierarchical clustering approach. Ticks represent patient censoring (patients leaving the study before completion). The top (green) line corresponds to patients with anomaly scores in roughly the top quartile of the population.

Clinical Variable	HR	P Value	95% CI
Age≥65 years	1.82	0.041	1.02-3.24
Female Gender	0.69	0.261	0.37-1.31
Current Smoker	1.05	0.866	0.59-1.87
Hypertension	1.44	0.257	0.77-2.68
Diabetes Mellitus	1.95	0.072	0.94-4.04
Hyperlipidemia	1.00	0.994	0.55-1.82
History of COPD	1.05	0.933	0.37-2.92
History of CHD	1.10	0.994	0.37-2.92
Previous MI	1.17	0.630	0.62-2.22
Previous angina	0.94	0.842	0.53-1.68
ST depression>0.5mm	1.13	0.675	0.64-2.01
Index diagnosis of MI	1.42	0.134	0.90-2.26

Table 3: Univariate association of existing clinical risk variables with major adverse cardiac events over a 90 day period following acute coronary syndrome.

believe that our data provides strong support for the ability of unsupervised risk stratification to add information beyond these existing metrics.

ECG Variable	HR	P Value	95% CI
HRV	1.56	0.128	0.88-2.77
HRT	1.64	0.013	1.11-2.42
DC	1.77	0.002	1.23-2.54

Table 4: Univariate association of existing ECG risk variables with major adverse cardiac events over a 90 day period following acute coronary syndrome.

	Age	Fem	Smo	Hpt	Dia	Lip	COPD	CHD	MI	Ang	ST	Ind
One-Class SVM	-0.07	0.02	-0.03	-0.08	-0.06	0.03	-0.03	0.03	-0.07	0.01	0.06	-0.02
3-Nearest Neighbor	0.11	0.00	-0.02	0.05	0.03	-0.05	0.04	-0.09	0.08	0.04	0.01	0.02
5-Nearest Neighbor	0.12	0.01	-0.03	0.05	0.05	-0.04	0.04	-0.10	0.09	0.05	0.02	0.02
7-Nearest Neighbor	0.11	0.00	-0.03	0.05	0.06	-0.04	0.04	-0.10	0.09	0.06	0.02	0.02
9-Nearest Neighbor	0.11	0.00	-0.02	0.05	0.06	-0.04	0.05	-0.10	0.09	0.07	0.01	0.02
Hierarchical Clustering	0.16	0.03	-0.04	0.05	0.08	-0.05	0.05	-0.08	0.04	0.00	0.03	0.04

Table 5: Correlation of high risk predictions with clinical risk variables (Age=age$\geq$65, Fem=female gender, Smo=current smoker, Hpt=hypertension, Dia=diabetes mellitus, Lip=hyperlipidemia, COPD=history of COPD, CHD=history of CHD, MI=previous MI, Ang=previous angina, ST=ST depression$>$0.5mm, Ind=Index diagnosis of MI).

	HRV	HRT	DC
One-Class SVM	-0.14	-0.01	-0.09
3-Nearest Neighbor	0.16	0.00	0.02
5-Nearest Neighbor	0.16	0.01	0.03
7-Nearest Neighbor	0.15	0.01	0.03
9-Nearest Neighbor	0.17	0.01	0.04
Hierarchical Clustering	0.20	0.06	0.08

Table 6: Correlation of high risk predictions with ECG risk variables.

6.2 Multivariate Results

We measured the correlation between the predictions of the unsupervised symbolic mismatch-based approaches and both the clinical and ECG risk variables. These results are shown in Tables 5 and 6. All of the unsupervised symbolic mismatch-based approaches had low correlation with both the clinical and ECG variables ($R \leq 0.2$).

Our results on multivariate analysis reflect this low correlation between the symbolic mismatch-based approaches and existing clinical and ECG risk variables. On multivariate analysis, both the nearest neighbor approach and the clustering approach were independent predictors of adverse outcomes (Table 7). In our study, the nearest neighbor approach (for $k > 3$) had the highest hazard ratio on both univariate and multivariate analysis. Both the nearest neighbor and clustering approaches predicted patients with an approximately two-fold increased risk of adverse outcomes. This increased risk did not change much even after adjusting for other clinical and ECG risk variables. While the classification approach did not perform as well, we note that this result may potentially be improved with availability of a larger data set to learn an enclosing boundary and by only using data from patients known to remain event-free on follow-up.

Method	Adjusted HR	P Value	95% CI
One-Class SVM	1.32	0.074	0.97-1.79
3-Nearest Neighbor	1.88	0.042	1.02-3.46
5-Nearest Neighbor	2.07	0.018	1.13-3.79
7-Nearest Neighbor	2.25	0.008	1.23-4.11
9-Nearest Neighbor	2.04	0.021	1.11-3.73
Hierarchical Clustering	1.86	0.042	1.02-3.46

Table 7: Multivariate association of high risk predictions from different approaches using symbolic mismatch with major adverse cardiac events over a 90 day period following acute coronary syndrome. Multivariate results adjusted for age$\geq$65 years, gender, smoking history, hypertension, diabetes mellitus, hyperlipidemia, history of COPD, history of CHD, previous MI, previous angina, ST depression on admission, index diagnosis of MI, HRV-LF/HF, HRT and DC.

Method	HR	P Value	95% CI
One-Class SVM	1.36	0.038	1.01-1.79
3-Nearest Neighbor	1.74	0.069	0.96-3.16
5-Nearest Neighbor	1.57	0.142	0.86-2.88
7-Nearest Neighbor	1.73	0.071	0.95-3.14
9-Nearest Neighbor	1.89	0.034	1.05-3.41
Hierarchical Clustering	1.19	0.563	0.67-2.12

Table 8: Univariate association of high risk predictions without the use of spectrum clipping. None of the approaches showed a statistically significant association with the study endpoint in any of the multivariate models including other clinical risk variables when spectrum clipping was not used.

Method	AUROC (Model 1)	AUROC (Model 2)
One-Class SVM	0.683	0.705
3-Nearest Neighbor	0.683	0.713
5-Nearest Neighbor	0.683	0.721
7-Nearest Neighbor	0.683	0.725
9-Nearest Neighbor	0.683	0.719
Hierarchical Clustering	0.683	0.711

Table 9: Improvement in discrimination when information obtained through unsupervised risk stratification is added to multivariate models containing age$\geq$65 years, gender, smoking history, hypertension, diabetes mellitus, hyperlipidemia, history of COPD, history of CHD, previous MI, previous angina, ST depression on admission, index diagnosis of MI, HRV-LF/HF, HRT and DC (Model A: existing risk variables, Model B: existing risk variables combined with unsupervised risk stratification).

Consistent with the univariate case above, we consider these findings to be encouraging. Our data suggests that the information available through unsupervised risk stratification based on sym-

bolic mismatch is generally independent of the information available through other specialized metrics. Moreover, our approach can potentially be used in a synergistic manner with these other metrics to improve risk stratification. For example, our study found that nearest neighbor-based risk stratification using symbolic mismatch can identify individuals who are at a two- to three-fold increased risk of adverse outcomes, even after adjusting for an extensive set of existing risk variables. This provides strong support for the potential ability of our research to complement present approaches to prognosticate cardiac patients. We hypothesize that these results are largely due to our focus on capturing information that is quite distinct from existing metrics. In particular, both our approach of risk stratifying patients within an unsupervised anomaly detection framework, and our focus on exploiting large volumes of long-term ECG data that is not well-suited for human analysis, predispose to capturing information that is clinically useful but not reflected in current metrics.

To quantify this effect better, we also studied how the area under the receiver operating characteristic curve (AUROC) changed for multivariate models constructed with and without the use of information obtained through unsupervised risk stratification. Table 9 presents the results obtained for this experiment. For each of the unsupervised risk stratification approaches, the addition of the information produced by these methods increased the ability of models developed using existing risk variables to discriminate between high and low risk patients. Consistent with the earlier results, this improvement was greatest for the 7-nearest neighbor approach. The results here provide further support for the information provided by our methods being potentially complementary to that captured by current risk variables.

6.3 Effect of Spectrum Clipping

We also investigated the effect of spectrum clipping on the performance of our different risk stratification approaches. Table 8 presents the associations when spectrum clipping was not used. For all three methods, performance was improved when spectrum clipping was used. We note that while our motivation for using spectrum clipping was largely to broaden the applicability of symbolic mismatch to kernel-based methods, the ability of spectrum clipping to reduce noise provided a positive effect for all methods.

6.4 Runtime Performance

Figure 6 presents a histogram of the number of heart beats in each patient's long-term ECG signal over the first 24 hours following admission. The median number of beats per patient was 99,581, with an interquartile range of 89,051 to 110,337. The minimum number of beats was 45,330 while the maximum was 161,696.

Figure 7 presents similar information for the number of symbols obtained per patient through the clustering process described in Section 3.1. The median number of symbols per patient was 66, with an interquartile range of 37 to 114. The minimum number of symbols was 1 while the maximum was 284.

On a dual-core Intel Pentium 4 3.06 GHz platform with 4GB RAM running MATLAB R2007a with Ubuntu 9.10 the symbolization of each patient's data (24 hours of ECG sampled at 128 Hz) took around 10 minutes. Roughly speaking, the use of symbolization compressed around 100,000 beats per patient to below 70 symbols (i.e., a reduction by a factor of just under 1,500). The corresponding improvement in the runtime of comparing long-term ECG signals was quadratic in this reduction, since (roughly speaking) instead of 100,000-by-100,000 comparisons of heart beats in the original

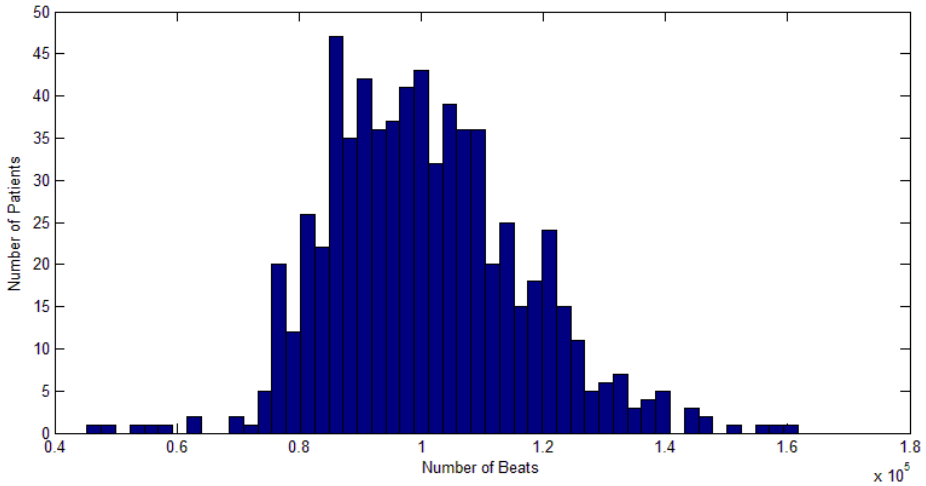

Figure 6: Histogram of the heart beats per patient over 24 hours (x-axis scale: $\times 10^5$).

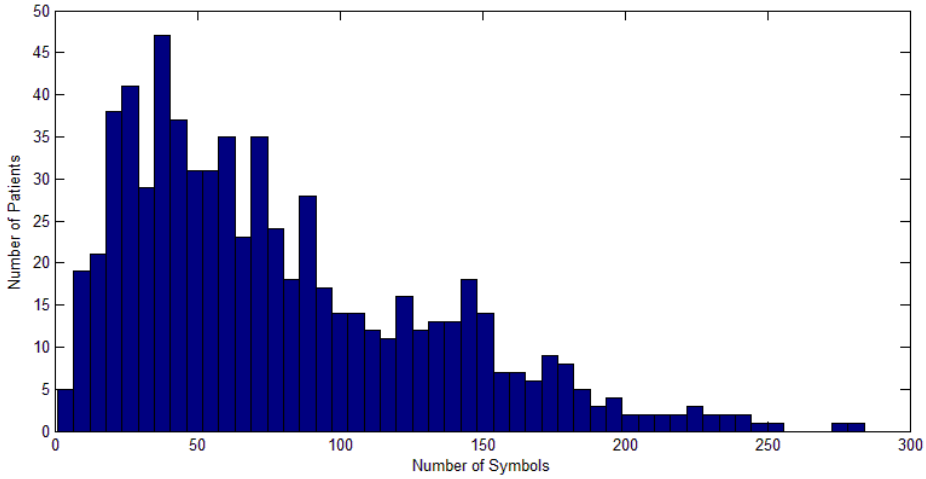

Figure 7: Histogram of the number of symbols per patient.

signals, only 70-by-70 symbol centroid and symbol frequency comparisons were necessary. The overall runtime complexity of our analysis was therefore $O(n^2\theta^2l^2) + O(n\theta ml^2)$, where the left term corresponds to the runtime of finding anomalies using symbolized data and the right term corresponds to the runtime of creating symbolic representations of the original ECG signals. We denote the number of patients by n, the maximum number of symbols for any patient by θ, the maximum number of heart beats for any patient by m, and the maximum length of any heart beat (in samples) by l. The left term of the runtime above is quadratic in the number of patients (since all pairs of patients are compared to find anomalies), the number of symbols (since all pairs of symbols are compared for any pair of patients), and the length of symbol centroids (since DTW compares all the samples for each pair of symbol centroids). The right term of the runtime above is linear in the number of patients (since each patient's data is symbolized once), with θm corresponding to the

time taken to make a pass through all m beats in each of the θ iterations of Max-Min clustering, and l^2 being the cost of using DTW to compare heart beats. While the computational cost of clustering is significant, it leads to a θ^2 factor in the runtime associated with finding anomalies using symbolic mismatch rather than an m^2 factor (where m is much larger than θ as shown by our results). The use of symbolization therefore represents one of the key sources of speedup in our study, reducing the runtime from $O(n^2m^2l^2)$ without symbolization to $O(n^2\theta^2l^2)+O(n\theta ml^2)$ with symbolization. We note that while other sources of runtime improvement are also possible (e.g., by addressing the quadratic runtime of DTW or by avoiding comparisons between all pairs of patients in the population for anomaly detection), the values of n and l are both much smaller, and therefore represent smaller gains, than the m^2 factor reduced by symbolization.

7. Related Work

Previous work on comparing time-series can be divided into two broad classes: methods to compare signals based on their raw samples, and methods to compare signals by extracting features from the data.

Most previous work on comparing signals in terms of their raw samples, including metrics such as dynamic time warping (Keogh and Pazzani, 2001; Keogh and Ratanamahatana, 2005), longest common subsequence (Vlachos et al., 2002), edit distance with real penalty (Cheng and Ng, 2004), sequence weighted alignment (Morse and Patel, 2007), spatial assembling distance (Chen et al., 2007), focuses on relatively short time-series. This is due to the runtime of these methods (quadratic for many methods) and the need to reason in terms of the frequency and dynamics of higher-level signal constructs (as opposed to individual samples) when studying systems over long periods. These existing methods do not, therefore, directly focus on comparing very long signals (e.g., tens of millions of samples).

In contrast to this, the vast majority of prior research on comparing long-term time-series focuses on extracting specific features from long-term signals and quantifying the differences between these features. For example, in the context of cardiovascular disease, long-term ECG is often reduced to features (e.g., mean heart rate or heart rate variability) and compared in terms of these features. These approaches, unlike our symbolic mismatch based approaches, draw upon significant *a priori* knowledge. Our belief was that for applications like risk stratifying patients for major cardiac events, focusing on a set of specialized features leads to important information being potentially missed. In our work, we focus instead on developing an approach that avoids use of significant *a priori* knowledge by comparing the raw morphology of long-term time-series. We propose doing this in a computationally efficient and systematic way through symbolization. While this use of symbolization represents a lossy compression of the original signal, the underlying process of quantifying differences between long-term time-series remains grounded in the comparison of raw morphology.

The process of symbolization maps the problem of comparing long-term time-series into the domain of sequence comparison. There is an extensive body of prior work focusing on the comparison of sequential or string data. Algorithms based on measuring the edit distance between strings are widely used in disciplines such as computational biology (Jones and Pevzner, 2004; Durbin et al., 1998), but are typically restricted to comparisons of short sequences because of their computational complexity. More closely related to our research is previous work on the use of profile hidden Markov models (Krogh, 1994; Jaakkola et al., 1999) to optimize recognition of binary labeled se-

quences. This work focuses on learning the parameters of a hidden Markov model that can represent approximations of sequences and can be used to score other sequences. Generally, this approach requires large amounts of data or sophisticated priors to train the hidden Markov models. Computing forward and backward probabilities from the Baum-Welch algorithm is also very computationally intensive. Subsequent research in this area focuses on mismatch tree-based kernels (Leslie et al., 2003), which use the mismatch tree data structure (Eskin and Pevzner, 2002) to quantify the difference between two sequences based on the approximate occurrence of fixed length subsequences within them. Similar to this approach is work on using a "bag of motifs" representation (Ben-Hur and Brutlag, 2006), which provides a more flexible representation than fixed length subsequences but usually requires prior knowledge of motifs in the data (Ben-Hur and Brutlag, 2006).

In contrast to these efforts, we use a simple, computationally efficient approach to compare symbolic sequences without prior knowledge. Most importantly, our approach helps address the scenario where symbolizing long-term time-series in a patient-specific manner leads to the symbolic sequences for patients being drawn from different alphabets (Syed et al., 2010). In this case, hidden Markov models, mismatch trees or a "bag of motifs" approach trained on one patient cannot be easily used to score the sequences for other patients. Instead, any comparative approach must maintain a hard or soft registration of symbols across individuals. Symbolic mismatch addresses this scenario and complements existing work on sequence comparison through a simple, computationally efficient measure that quantifies differences across patients while retaining information on how the symbols for these patients differ.

Finally, we also distinguish our work from earlier efforts to risk stratify patients using long-term data. In particular, we supplement our symbolic mismatch kernel with the idea of detecting high risk patients as those individuals in the population with unusual long-term time-series. For example, in the context of cardiovascular disease, techniques such as heart rate variability (Malik et al., 1996), heart rate turbulence (Barthel et al., 2003), and t-wave alternans (Smith et al., 1988; Rosenbaum et al., 1994) have all been shown to be useful in risk stratifying patients at risk for future cardiovascular events following acute coronary syndromes. The focus of these methods is to calculate a particular pre-defined feature from the raw ECG signal, and to use it to rank patients along a risk continuum. Our approach, focusing on detecting patients with high symbolic mismatch relative to other patients in the population, is orthogonal to the use of specialized high risk features along two important dimensions. First, it does not require the presence of significant prior knowledge. For the cardiovascular care, we only assume that ECG signals from patients who are at high risk differ from those of the rest of the population. There are no specific assumptions about the nature of these differences. Second, the ability to partition patients into groups with similar ECG characteristics and potentially common risk profiles potentially allows for a more fine-grained understanding of a how a patient's future health may evolve over time. Matching patients to past cases with similar ECG signals could lead to more accurate assignments of risk scores for particular events such as death and recurring heart attacks.

8. Discussion

In this paper, we proposed using symbolic mismatch to quantify differences in long-term physiological time-series. Our approach uses a symbolic transformation to measure changes in the morphology and frequency of prototypical functional units observed over long periods in two signals.

In addition to proposing symbolic mismatch, which avoids feature extraction and deals with inter-patient differences in a parameter-less way, we also explored the hypothesis that high risk patients in a population can be identified as individuals with anomalous long-term signals. We developed multiple comparative approaches to detect such patients, and evaluated these methods in a real-world application of risk stratification for major adverse cardiac events following acute coronary syndrome. Our results suggest that symbolic mismatch-based comparative approaches may have clinical utility in identifying high risk patients, and can provide information that is complementary to existing clinical risk variables.

In particular, we note that the hazard ratios we report are typically considered clinically meaningful. Risk stratification following ACS is an extremely challenging goal. In a different study of 118 variables in 15,000 post-ACS patients with 90 day follow-up similar to our population, Newby et al. (2003) did not find any variables with a hazard ratio greater than 2.00. We observed a similar result in our patient population, where all of the existing clinical and ECG risk variables had a hazard ratio less than 2.00. In contrast to this, our nearest neighbor-based approach achieved a hazard ratio of 2.28, even after being adjusted for existing risk measures. We believe that these results provide strong support for the potential role of our research in improving the management of patients post-ACS.

We envision our techniques being primarily useful in their ability to enrich models for cardiovascular risk stratification. In other words, we expect the risk scores generated by our methods to serve as features that can be combined with other features based on specialized knowledge while assessing the overall health of patients. While we dichotomized the results of all of our methods for evaluation consistent with the way most new cardiovascular risk metrics are validated, we believe that the best use of this information is in its original continuous form to provide a finer grained distinction between high and low risk patients. We further believe that the eventual use case of our tools will be to assess individual patients that present at different times as anomalies relative to a continuously increasing data set of patients seen previously. Aspects of our research, such as symbolic mismatch, may also have a role in a supervised setting, as we discuss later in this section.

In the context of cardiovascular disease, techniques such as heart rate variability, heart rate turbulence, T wave alternans, and morphologic variability have all been shown to be useful in risk stratifying patients at risk for future cardiovascular events following acute coronary syndromes. The focus of these methods is to calculate a specific pre-defined feature from the raw ECG signal, and to use it to rank patients along a risk continuum. Our approach, focusing on detecting patients with high symbolic mismatch relative to other patients in the population, is orthogonal (and perhaps complementary) to the use of specialized high risk features. First, it does not require the presence of significant prior knowledge. For the cardiovascular care, we only assume that ECG signals from patients who are at high risk differ from those of the rest of the population. There are no specific assumptions about the nature of these differences. Second, the ability to partition patients into groups with similar ECG characteristics and potentially common risk profiles potentially allows for a more fine-grained understanding of a how a patient's future health may evolve over time. Matching patients to past cases with similar ECG signals could lead to more accurate assignments of risk scores for particular events such as death and recurring heart attacks.

We conclude with some limitations of our work. While our decision to compare the morphology and frequency of prototypical functional units leads to a measure that is computationally efficient on large volumes of data, this process does not capture information related to the dynamics of these prototypical units or in specific sequences of symbols. We also observe that all three of the com-

parative approaches investigated in our study focus only on identifying patients who are anomalies. Although we believe that symbolic mismatch may have further use in supervised learning, the size of our patient population and the small number of adverse cardiac outcomes over the 90 day follow-up meant that dividing the patients into separate training and testing groups would make it challenging to learn models that generalized well. This hypothesis, that is, of symbolic mismatch being useful in the setting of supervised learning, therefore needs to be evaluated more fully on larger patient populations. Finally, we note that we did not have echocardiographic data for patients in the DISPERSE2 trial. As a result, we did not include a comparison in our study to metrics such as left ventricular ejection fraction (LVEF). We believe that our research warrants further investigation on larger data sets, with a more comprehensive set of existing clinical metrics, and longer follow-ups in the future.

Acknowledgments

We thank GE Healthcare and Georg Schmidt for sharing the libRASCH toolbox to measure HRT and DC values. This research was supported by the National Science Foundation CAREER award 1054419.

References

AAMI. Recommended practice for testing and reporting performance results of ventricular arrhythmia detection algorithms. 1987.

AAMI. Testing and reporting performance results of cardiac rhythm and ST segment measurement algorithms. 1998.

J.S. Alpert. Risk stratification for cardiac events after acute ST elevation myocardial infarction. *UpToDate*, 2010.

J.J. Bailey, A.S. Berson, H. Handelsman, and M. Hodges. Utility of current risk stratification tests for predicting major arrhythmic events after myocardial infarction. *Journal of the American College of Cardiology*, 38(7):1902–1911, 2001.

P. Barthel, R. Schneider, A. Bauer, K. Ulm, C. Schmitt, A. Schomig, and G. Schmidt. Risk stratification after acute myocardial infarction by heart rate turbulence. *Circulation*, 108(10):1221–1226, 2003.

A. Bauer, J.W. Kantelhardt, P. Barthel, R. Schneider, T. Makikallio, K. Ulm, K. Hnatkova, A. Schomig, H. Huikuri, A. Bunde, et al. Deceleration capacity of heart rate as a predictor of mortality after myocardial infarction: cohort study. *The Lancet*, 367(9523):1674–1681, 2006.

A. Ben-Hur and D. Brutlag. Sequence motifs: highly predictive features of protein function. *Feature Extraction*, pages 625–645, 2006.

J.A Breall and M. Simons. Risk stratification after unstable angina or non-ST elevation myocardial infarction. *UpToDate*, 2010.

C.P. Cannon, S. Husted, R.A. Harrington, B.M. Scirica, H. Emanuelsson, G. Peters, and R.F. Storey. Safety, Tolerability, and Initial Efficacy of AZD6140, the First Reversible Oral Adenosine Diphosphate Receptor Antagonist, Compared With Clopidogrel, in Patients With Non–ST-Segment Elevation Acute Coronary Syndrome Primary Results of the DISPERSE-2 Trial. *Journal of the American College of Cardiology*, 50(19):1844–1851, 2007.

S.H. Chang, F.H. Cheng, W. Hsu, and G. Wu. Fast algorithm for point pattern matching: invariant to translations, rotations and scale changes. *Pattern Recognition*, 30(2):311–320, 1997.

Y. Chen, MA Nascimento, B.C. Ooi, and AKH Tung. Spade: On shape-based pattern detection in streaming time series. In *IEEE International Conference on Data Engineering*, pages 786–795, 2007.

Y. Chen, E.K. Garcia, M.R. Gupta, A. Rahimi, and L. Cazzanti. Similarity-based classification: Concepts and algorithms. *Journal of Machine Learning Research*, 10:747–776, 2009a.

Y. Chen, MR. Gupta, and B. Recht. Learning kernels from indefinite similarities. In *International Conference on Machine Learning*, pages 145–152, New York, NY, USA, 2009b. ACM. ISBN 978-1-60558-516-1.

L. Cheng and R.T. Ng. On The marriage of LP-norms and edit distance. In *Very Large Data Bases*, pages 792–803, 2004.

W.W. Cohen and J. Richman. Learning to match and cluster large high-dimensional data sets for data integration. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 475–480. ACM New York, NY, USA, 2002.

T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.

D.R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 187–220, 1972.

P. de Chazal, M. O'Dwyer, and R.B. Reilly. Automatic classification of heartbeats using ECG morphology and heartbeat interval features. In *IEEE Transactions on Biomedical Engineering*, pages 1196–1206, 2004.

C.H. Ding, X. He, H. Zha, M. Gu, and H. Simon. A min-max cut algorithm for graph partitioning and data clustering. In *IEEE International Conference on Data Mining*, pages 107–114, 2001.

R. Durbin, S.R. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1998.

B. Efron. Logistic regression, survival analysis, and the Kaplan-Meier curve. *Journal of the American Statistical Association*, 83(402):414–425, 1988.

E. Eskin and P.A. Pevzner. Finding composite regulatory patterns in DNA sequences. *Bioinformatics*, 18:354–363, 2002.

E. Eskin, A. Arnold, M. Prerau, L. Portnoy, and S. Stolfo. A geometric framework for unsupervised anomaly detection. *Applications of Data Mining in Computer Security*, pages 77–200, 2002.

B. Goldstein. Intensive care unit ECG monitoring. *Cardiac Electrophysiology Review*, 1(3):308–310, 1997. ISSN 1385-2264.

P.S. Hamilton and W.J. Tompkins. Quantitative investigation of QRS detection rules using the MIT/BIH arrhythmia database. *IEEE Transactions on Biomedical Engineering*, 33(12):1157–1165, 1986.

T. Jaakkola, M. Diekhans, and D. Haussler. Using the Fisher kernel method to detect remote protein homologies. In *Proceedings of the International Conference on Intelligent Systems for Molecular Biology*, pages 149–158, 1999.

N.C. Jones and P. Pevzner. *An Introduction to Bioinformatics Algorithms*. the MIT Press, 2004.

E. Keogh and C.A. Ratanamahatana. Exact indexing of dynamic time warping. *Knowledge and Information Systems*, 7(3):358–386, 2005.

E.J. Keogh and M.J. Pazzani. Derivative dynamic time warping. In *SIAM international Conference on Data Mining*. Citeseer, 2001.

R.E. Kleiger, P.K. Stein, and J.T. Bigger Jr. Heart rate variability: measurement and clinical utility. *Annals of Noninvasive Electrocardiology*, 10(1):88–101, 2005. ISSN 1542-474X.

A. Krogh. Hidden Markov models for labeled sequences. In *International Conference on Pattern Recognition*, pages 140–144. IEEE Computer Society Press, 1994.

C. Leslie, E. Eskin, J. Weston, and W.S. Noble. Mismatch string kernels for SVM protein classification. *Advances in Neural Information Processing Systems*, pages 1441–1448, 2003.

L.S. Lilly. Pathophysiology of Heart Failure. 2007.

G. Lopera and A.B. Curtis. Risk stratification for sudden cardiac death: current approaches and predictive value. *Current Cardiology Reviews*, 5:56–64, 2009.

J. Mackay, G.A. Mensah, S. Mendis, and K. Greenlund. *The Atlas of Heart Disease and Stroke*. World Health Organization, 2004.

M. Malik, J.T. Bigger, A.J. Camm, R.E. Kleiger, A. Malliani, A.J. Moss, and P.J. Schwartz. Heart rate variability: Standards of measurement, physiological interpretation, and clinical use. *European Heart Journal*, 17(3):354–381, 1996.

M.D. Morse and J.M. Patel. An efficient and accurate method for evaluating time series similarity. In *ACM SIGMOD International Conference on Management of Data*, pages 569–580. ACM, 2007.

CS Myers and LR Rabiner. A comparative study of several dynamic time-warping algorithms for connected word recognition. *Bell Systems Technical Journal*, 60(7):1389–1409, 1981.

B.D. Nearing and R.L. Verrier. Modified moving average analysis of T-wave alternans to predict ventricular fibrillation with high accuracy. *Journal of Applied Physiology*, 92(2):541–549, 2002.

L.K. Newby, M.V. Bhapkar, H.D. White, E.J. Topol, F.C. Dougherty, R.A. Harrington, M.C. Smith, L.F. Asarch, R.M. Califf, et al. Predictors of 90-day outcome in patients stabilized after acute coronary syndromes. *European Heart Journal*, 24(2):172–181, 2003.

Physionet. MIT-BIH Arrhythmia Database. 2005.

D.S. Rosenbaum, L.E. Jackson, J.M. Smith, H. Garan, J.N. Ruskin, and R.J. Cohen. Electrical alternans and vulnerability to ventricular arrhythmias. *New England Journal of Medicine*, 330 (4):235–241, 1994.

B. Scholkopf and A.J. Smola. *Learning with Kernels*. MIT Press, 2002.

B. Scholkopf, J.C. Platt, J. Shawe-Taylor, A.J. Smola, and R.C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471, 2001.

M.G. Shlipak, J.H. Ix, K. Bibbins-Domingo, F. Lin, and M.A. Whooley. Biomarkers to predict recurrent cardiovascular disease: the Heart and Soul Study. *The American Journal of Medicine*, 121(1):50–57, 2008.

J.M. Smith, E.A. Clancy, C.R. Valeri, J.N. Ruskin, and R.J. Cohen. Electrical alternans and cardiac electrical instability. *Circulation*, 77(1):110–121, 1988.

Z. Syed, J. Guttag, and C. Stultz. Clustering and symbolic analysis of cardiovascular signals: Discovery and visualization of medically relevant patterns in long-term data using limited prior knowledge. *EURASIP Journal on Advances in Signal Processing*, 2007:14, 2007.

Z. Syed, C. Stultz, M. Kellis, P. Indyk, and J. Guttag. Motif discovery in physiological datasets: a methodology for inferring predictive elements. *ACM Transactions on Knowledge Discovery from Data*, 4(1):1–23, 2010.

B. Vanschoenwinkel and B. Manderick. Appropriate kernel functions for support vector machine learning with sequences of symbolic data. *Lecture Notes in Computer Science*, 3635/2005:256–280, 2005.

M. Vlachos, D. Gunopoulos, and G. Kollios. Discovering similar multidimensional trajectories. In *IEEE International Conference on Data Engineering*, pages 673–684. Published by the IEEE Computer Society, 2002.

J.H. Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963.

World Health Organization. Cardiovascular diseases (CVDs) fact sheet, 2009.

G. Wu, E.Y. Chang, and Z. Zhang. An analysis of transformation on non-positive semidefinite similarity matrix for kernel machines. Technical report, University of California, Santa Barbara, 2005.

W. Zong, G.B. Moody, and D. Jiang. A robust open-source algorithm to detect onset and duration of QRS complexes. *Computers in Cardiology*, pages 737–740, 2003.