Computational Cardiology:
Improving Markers and Models to Stratify Patients with Heart Disease

by

Chih-Chun Chia

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Electrical Engineering: Systems)
in the University of Michigan
2014

Doctoral Committee:

Professor Zeeshan H. Syed, Co-Chair
Professor Satinder Singh Baveja, Co-Chair
Professor Laura K. Balzano
Professor James M. Blum
Professor Emily Mower Provost
Professor Clayton D. Scott

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

CHAPTER I

Introduction

In this thesis, we explore the question of how the burden of coronary artery disease can be significantly reduced through advances in computation. The general problem statement for our work is to develop novel approaches that can be applied to large physiological datasets to identify new markers of cardiovascular disease and to improve models to predict adverse clinical outcomes. This research proposes techniques related to machine learning, signal processing, and algorithm design; and develops these ideas within the context of cardiovascular physiology and the clinical use-case for important medical challenges.

1.1 Motivation

Cardiovascular disease is the leading cause of death in the United States, claiming over 830,000 lives each year (34% of all deaths, or roughly one death every 38 seconds) [63]. More than 151,000 of these deaths take place in patients under the age of 65, and a third occur before the age of 75. Moreover, with the aging of the U.S. population and the demographic changes projected for the year 2020 and beyond, it is expected that cardiovascular disease will continue to be a major healthcare challenge in the years to come. The situation is similar in other parts of the world. Estimates show that the number of cardiovascular deaths (CVDs) will increase both in developed countries (6 million vs. 5 million) and developing countries (19 million vs. 9 million) between the years 2000 and 2020 [85]. These statistics are particularly grim for the developing world, where more than 40% of all the deaths by the end of this decade are expected to be due to cardiovascular disease [85].

One of the difficulties of dealing with cardiovascular disease, and coronary heart disease in particular, is that despite the availability of different treatment options the disease burden remains unacceptably high because of an inability to match patients to treatments that are most appropriate for them individually. One of the best examples of this situation is provided by implantable cardiac defibrillators (ICDs), which can be life-saving for patients who experience fatal arrhythmias (over 300,000 sudden cardiac deaths in the U.S. each year among patients with diagnosed coronary disease) [75, 17]. In most of these cases, the effects of the arrhythmia can be reversed if the victim is treated with an electrical shock within the first few minutes. However, existing decision-making methods fail to prescribe ICDs to the majority of patients who die [17]. Conversely, 90% of the patients who do currently receive an ICD do not receive any benefit from their device [17], resulting in an unnecessary risk to patients and unnecessary costs to the healthcare system.

1.2 Overview

With advances in recording and storage technologies, and with the increasing use of electronic health record (EHR) systems in hospitals and clinics, it is now possible to collect larger volumes of data than was previously thought possible. This increase has taken place both in terms of the amount of data that is recorded per patient, and in terms of the number of patients monitored. These data offer an enormous opportunity to advance cardiac care through new data-derived knowledge. Given the limitations of existing approaches grounded in *a priori* assumptions about cardiac disease, this knowledge is essential to supplementing current decision-making tools for patient care.

Our work explores the opportunity of improving cardiovascular decision-making in this context. Of particular interest in our work to develop and validate data-driven advances for cardiovascular care is the electrocardiogram (ECG). Despite the ECG being a signal that has been extensively studied for over a hundred years, it offers a prototypical example of how computation can allow for routinely collected clinical data to be used more meaningfully. Specifically, the ECG contains a wealth of information related to both the electrical activity of the heart, and the regulation of cardiac function by the autonomic nervous system. The extensive information provided by the ECG signal, as well as the simple, non-invasive and relatively inexpensive nature of ECG acquisition, have made ECG-based metrics attractive for cardiovascular risk stratification. There is also an opportunity to leverage the ECG for fundamental advances in medical knowledge and cardiovascular care because most existing research related to ECG-based risk stratification has focused on studying short (e.g., 30 second) snapshots of ECG. With improvements in our ability to collect ECG data over long periods and in a digital format conducive to computational analysis, our work explores the discovery of computational biomarkers from long-term ECG and their incorporation in broad clinical practice.

In addition to focusing on the ECG, we also address the question of how information in the new biomarkers discovered through our research can be integrated alongside existing clinical parameters to holistically stratify patients. In doing so, we address the question of how the information spanning a broad range of demographic, comorbidity, history, physical exam, medication, and procedural data can be leveraged to more comprehensively evaluate patients. In contrast to our efforts related to the ECG (where essentially the long nature of the recordings makes it difficult to effectively leverage information) in the setting of building multi-factorial models we address the question of how high-dimensional information can be used to improve patient care.

In summary, the two major directions that we explore in this thesis are (1) extracting novel features to measure subtle but persistent ECG changes that have high prognostic value; and (2) building personalized predictive models that combine information in both novel and established risk metrics to assess individual health. Our research in both these areas is synergistic; and creates the opportunity for substantially improving the diagnosis and management of cardiovascular patients.

1.3 Major Contributions and Results

We briefly list the major contributions of this thesis here. A detailed discussion of these contributions is deferred to the subsequent chapters. This section aims to provide a short overview of the major accomplishments of our research and how the different efforts can be viewed as synergistically integrating together for next-generation cardiovascular care. Some specific achievements arising from our work include:

- Improving on existing research in the area of morphologic variability (MV) by developing an adaptive downsampling-based approach that results in a four-fold improvement in computational efficiency while maintaining clinical discrimination.

- Developing a new ECG biomarker through the use of randomized hashing that studies information in short-term heart rate patterns to evaluate cardiac autonomic regulation. Evaluation on a cohort of over 3000 patients shows that our proposed marker is associated with a two-fold increased risk of cardiovascular mortality following acute coronary syndrome (ACS) even after adjusting for existing clinical markers.

- Exploring a novel 1.5-class learning paradigm for clinical models that is intended to address the issues of small datasets and class imbalance affecting clinical applications. Our use of an approach that leverages the best properties of both supervised and unsupervised learning significantly improves the discrimination for adverse mortality and morbidity endpoints relative to different conventional algorithms across a range of clinical applications.

- Proposing and demonstrating the hypothesis that information in the atrial component of the ECG can aid in the stratification of patients for post-operative atrial fibrillation (PAF). Using a new source separation algorithm we showed on a cohort of 385 patients undergoing cardiac surgery that our approach can improve reclassification of patients by over 25% relative to the use of existing cardiovascular markers.

1.4 Organization

The remainder of this thesis is organized as follows. Chapter 2 reviews essential clinical background. Chapters 3 and 4 then present our research related to ECG-based stratification of cardiac patients. In particular, Chapter 3 describes our work to improve morphology-based markers through adaptive downsampling while Chapter 4 details our work to develop new rate-based markers through randomized hashing. This is followed by a discussion in Chapter 5 of our efforts to build improved models for clinical risk stratification using a novel 1.5-class learning paradigm. In Chapter 6, we supplement our efforts related to ventricular arrhythmias in the setting of acute coronary syndrome (Chapters 3 to 5) with a different application: stratifying patients for atrial fibrillation following cardiac surgery. Finally, we conclude in Chapter 7 with a summary of our work and a discussion of potential future research avenues arising from this thesis.

CHAPTER II

Background

In this chapter, we review background that may be helpful for following the contributions of this thesis. We start with a brief introduction to ACS and discuss existing approaches used to stratify patients post-ACS. This is followed by a more detailed review of one of these approaches for post-ACS stratification, i.e. ECG, which forms a major focus of our work. We focus, in particular, on reviewing different ECG-based metrics for evaluating both the health of the heart as well as the health of the nervous system regulating cardiac activity. These details provide useful context for the contributions presented subsequently in this thesis.

2.1 Post-Acute Coronary Syndrome Stratification

An ACS is a clinical term associated with a spectrum of disorders (i.e., myocardial infarction and unstable angina) where the blood supply to the heart is suddenly blocked. The most common symptom is unusual chest pain with a tightness around the chest. Other symptoms include diaphoresis, nausea and vomiting, as well as shortness of breath. In some cases (particularly patients with diabetes) these symptoms may be absent altogether corresponding to a silent heart attack.

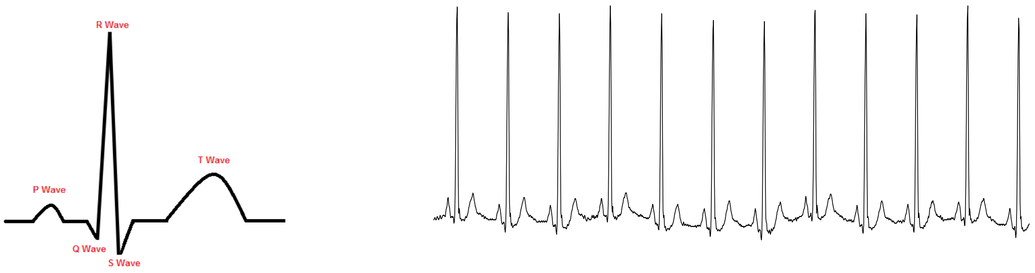
One of the challenges associated with ACS is that patients who survive the index event and receive treatment to open up the blood vessels supplying oxygen to the heart still remain at elevated risk of mortality and morbidity over follow-up. This is, in part, because ACS may cause permanent damage to heart tissue affecting the electrical and mechanical function of the heart, and leading to irregular heart rhythms (arrhythmias) that can be fatal.

Patients at risk of arrhythmias following ACS can derive substantial benefit from aggressive monitoring and therapy. As a result, post-ACS risk stratification is extremely important clinically for determining cardiac care.

Physicians use a variety of biomarkers to estimate patient risk and to match patients to treatments. These biomarkers are typically limited to information available through blood-based measurements of biochemical substrates (e.g. troponin I, C-reactive protein, and brain natriuretic peptide), or through imaging (e.g., left ventricular ejection fraction obtained through echocardiography) [59, 56]. In both these cases, the focus is on studying information that is present in instantaneous (i.e., 'snapshot') data, and where this information can be directly measured with limited or no computational aid. Despite these efforts, however, finding biomarkers that can accurately assess patient risk remains a challenge. For instance, while depressed left ventricular ejection fraction is commonly used to identify high risk patients following heart attacks, the absolute number of deaths is far greater among patients with relatively preserved left ventricular function [38].

2.2 Electrocardiogram (ECG)

The ECG is a continuous recording of the electrical activity of the heart muscle or myocardium [60]. At rest, each cardiac muscle cell maintains a voltage difference across its cell membrane. During depolarization (i.e., the 'firing' of the heart muscle), this voltage increases. 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



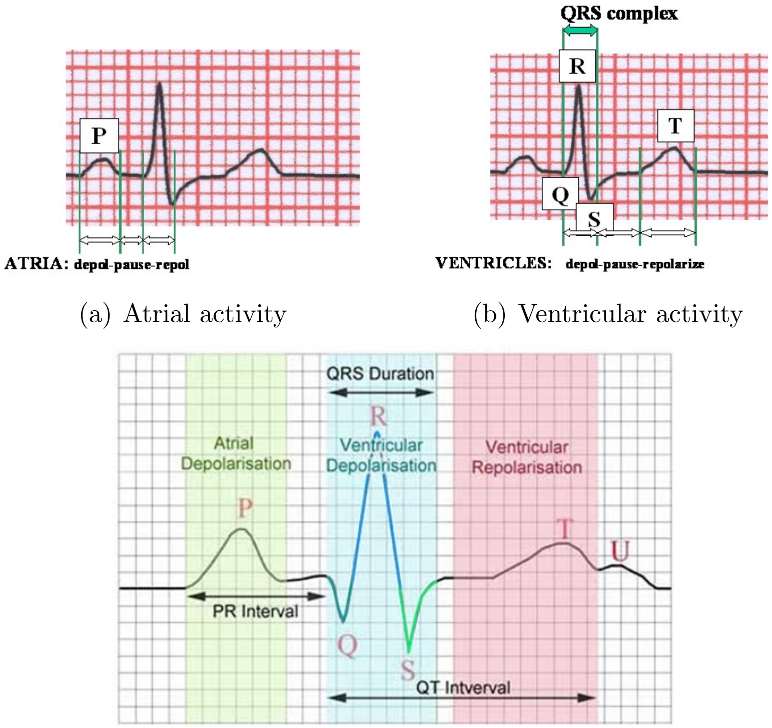
(a) Single ECG beat

(b) Continuous ECG tracing

Figure 2.1: (a) Schematic representation of the normal ECG for a single heart beat, and (b) example recording of ECG waveform.

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, leading to the ECG time-series.

Figure 2.1(a) presents a schematic representation of the normal ECG, while Figure 2.1(b) shows an example tracing of a continuous ECG time-series over a few seconds. We can see that 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, namely, the P-wave, the QRS-complex, and the T-wave. As shown in Figure 2.2(c) 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 smaller amplitude than the QRS-complex because repolarization takes longer than depolarization [60]. Figure. 2.2 shows the relationship between ECG signal (P, QRS, T-waves) and atrial/ventricular polarization.



(a) Atrial activity

(b) Ventricular activity

(c) ECG

Figure 2.2: ECG

2.3 ECG Metrics

2.3.1 Why ECG Metrics

With advances in recording and storage technologies, far larger volumes of time-series data are now collected from individual patients than was previously thought possible. This includes increases in the durations and sampling rates of recordings, as well as improvements in the ability to monitor patients in a variety of hospital and ambulatory settings. The large databases of physiological time-series resulting from this progress provide an opportunity to measure fundamentally new kinds of clinical information related to subtle phenomena occurring over long time scales. Given this opportunity, our research aims to address post-ACS risk stratification problem through novel biomarkers that are computationally derived from physiological time-series.

Of particular interest, in the setting of heart disease, is information available in long-term ECG data. The ECG provides a continuous assessment of the electrical activity of the heart, and is routinely collected from patients during hospitalization to determine heart rate and detect arrhythmias. The 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 over long periods. 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. Since the ECG is routinely collected from patients in a wide variety of clinical settings during patient hospitalization, computational biomarkers deriving from long-term ECG time-series can be incorporated broadly into clinical practice without the need for any new hardware or without creating any additional burden on patients or caregivers.

The ECG contains a wealth of information related to the structure of the heart (e.g., size of chambers, thickness of chamber walls, presence of scar tissue due to old infarcts), the function of the heart (e.g., rate of conduction, sequence in which different parts of the heart initiate impulses and contract, and stability of conduction), the influence of the autonomic nervous system on the heart (e.g., variation in heart rate), and the health of the coronary vasculature supplying blood to the heart (e.g., presence of ischemia) [48]. As a result, recent years have seen a growing interest in leveraging ECG signals to stratify patients.

2.3.2 Risk Variables Extracted from Long-term ECG-based Time-series

There is a substantial body of research focusing on ECG-based risk stratification of cardiac patients [68, 27, 101, 103, 100]. A number of different ECG metrics proposed for post-ACS risk stratification have shown promise. These include: (1) Heart rate variability (HRV) [79], which will be described in the next paragraph; (2) Heart rate turbulence (HRT) [11], which measures the return to equilibrium of the heart rate after a premature ventricular contraction (PVC); (3) Deceleration capacity (DC) [13], which measures cardiac responsiveness to vagal stimulation; (4) Severe autonomic failure (SAF) [12], which combines HRT and DC to more completely characterize risk; (5) Signal averaged ECG (SAECG) [90], which averages multiple ECG signals to remove interference and reveal small variations in the QRS-complex; (6) QT dispersion [37], which measures the variation between QT intervals across different leads of ECG signals; and (7) T-wave alternans (TWA) [18], which studies the morphology of the ECG signal for repolarization abnormalities. Despite the promise of these metrics, however, few of these metrics have been incorporated into clinical practice due to issues related to precision and recall, inconsistent findings across populations, and uncertain analytic performance.

Heart Rate Variability: HRV is the most widely used existing ECG marker. It studies variability in the length of normal (sinus) heart beats to assess cardiac autonomic function, i.e., the controlling influence of the nervous system on the heart. The intuition underlying HRV is that diminished variability in heart rate during normal (sinus) rhythm suggests impaired cardiac autonomic regulation. For example, the lack of variation in heart rate over extended periods suggests that the nervous system is not regulating cardiac activity to track differences in physiological and physical activity. Patients with decreased HRV are therefore believed to be at an increased risk of adverse outcomes, especially fatal arrhythmias, since the protective regulatory effects of the autonomic nervous system are similarly curtailed while dealing with

Table 2.1: Time- and frequency-domain measures of HRV (NN intervals = lengths of normal beats) proposed by the Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology. Each of these metrics represents a fairly simple approach to capture aggregate variability and loses information about specific structure in heart rate changes corresponding to patient risk.

Variable	Definition
SDNN	Standard deviation of NN intervals over 24-hour ECG
SDANN	Standard deviation of the average NN intervals for all 5-minute segments of 24-hour ECG
ASDNN	Average of the standard deviation of NN intervals for all 5-minute segments of 24-hour ECG
RMSSD	Square root of the mean squared differences of successive NN intervals
PNN50	Ratio derived by dividing the number of interval differences of successive NN intervals $>$ 50ms by the total number of NN intervals
HRVI	Total number of all NN intervals divided by the height of the histogram of all NN intervals measured on a discrete scale with bins of 7.8125ms
LFHF	Average ratio of the power in the frequency spectrum of 5-minute windows of the time series between 0.04-0.15 Hz and 0.15-0.4 Hz

abnormalities. Table 2.1 summarizes some of the different approaches proposed to measure HRV from long-term measurements of heart rate in ECG data.

2.4 Morphologic Variability (MV)

Unlike HRV, which focuses only the relative spacing of the R-waves (i.e., the spacing between beats), MV is a novel marker that was recently proposed as a means of leveraging information present in the entire ECG tracing [104]. The hypothesis underlying MV is that much of what is commonly perceived as noise in ECG data may contain subtle but useful information about the health of the heart. Specifically, it is believed that increased variability in the morphology of the ECG time-series is likely associated with a lack of consistency and repeatability in the electrical function of the heart. In other words, persistent fluctuations in the shape of the ECG waveform may be associated with electrical instability in the heart muscle predisposing patients to fatal arrhythmias. There is support for this theory in the literature. Studies have shown that in the presence of ischemia, the conducting system has multiple irregular islands of depressed myocardium with relatively long refractory times [32] that leads to discontinuous electrophysiological characteristics [65]. The overall effect of such minor conduction inhomogeneities is not well understood, but it is speculated that they correlate with myocardial electrical instability and have potentially predictive value for ventricular arrhythmias [14] or other adverse events.

The challenge in detecting this variability, however, is in being able to distinguish between shape deformations associated with pathological phenomena reflecting the health of the underlying heart muscle, and changes associated with artifacts that represent true noise. Making this distinction is difficult in short ECG recordings, but with the availability of long-term ECG time-series, pathological variations can be distinguished from true noise as structure that is persistent over long periods of time. In the remainder of this section, we describe our approach to measure MV from ECG time-series.

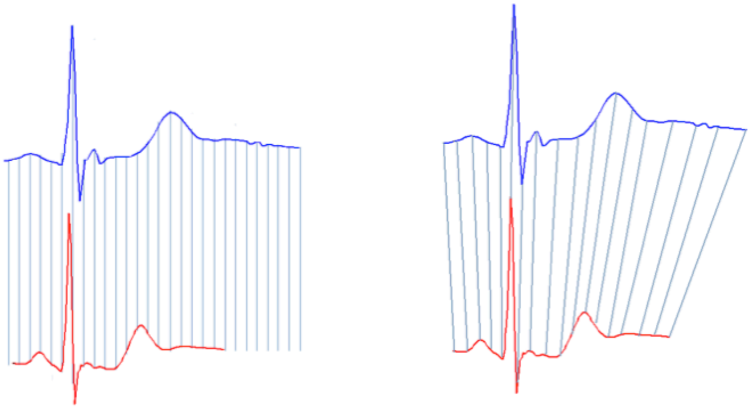


Figure 2.3: Comparison of time-warped shape deformations in ECG beats using DTW. In contrast to comparing activity that is time-aligned but not physiologically aligned (left), we use DTW in our study to relate similar parts of the ECG waveforms across beats in the presence of time skew.

For every pair of consecutively occurring beats in an ECG time-series, we quantify how the shapes of the beats differ using a variant of dynamic time warping (DTW, as in Figure 2.3). Given time-series $Q = q_1, \ldots, q_n$ and $C = c_1, \ldots, c_m$, DTW first constructs an n-by-m distance matrix where each entry (i, j) represents the distance $d(q_i, c_j)$. The l_2 norm is typically used to measure $d(q_i, c_j)$. DTW then finds the minimum cost path $W = w_1, \ldots, w_k, \ldots, w_K$ through this distance matrix where $w_k = (i_k, j_k)$ relates the i_k-th sample of Q to the j_k-th sample of C. The minimum cost path has the cost:

$$\sum_{k=1}^{K} d(q_{i_k}, c_{j_k})$$

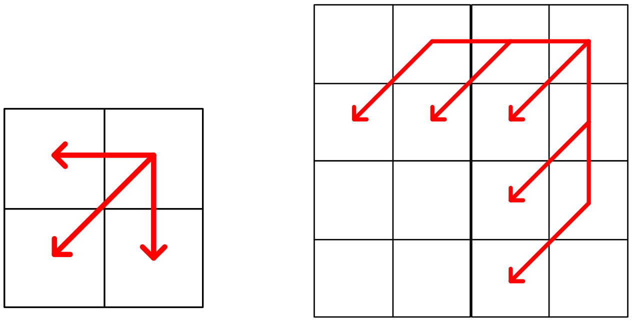
and is subject to several constraints, including boundary conditions, continuity, and monotonicity[16]. This optimal path can be found efficiently using dynamic programming with the following recurrence:

$$\gamma(i, j) = d(q_i, c_j) + \min \begin{cases} \gamma(i-1, j-1) \\ \gamma(i-1, j) \\ \gamma(i, j-1) \end{cases}$$

where $\gamma(i, j)$ is the cumulative distance of the path from the start to cells (i, j). From simple observation, $DTW(Q, C) = \gamma(n, m)$ and the time and space complexity of this method is $O(nm)$.

We restrict the local range of the alignment path in the vicinity of a point to prevent biologically implausible alignments of large parts of one beat with small parts of another. For example, for an entry (i, j) in the distance matrix, we only allow valid paths passing through $(i-1, j-1)$, $(i-1, j-2)$, $(i-2, j-1)$, $(i-1, j-3)$ and $(i-3, j-1)$. This is an adaptation of the Type III and Type IV local continuity constraints proposed by Myers et al. [76] and ensures that there are no long horizontal or vertical edges along the optimal path through the distance matrix, corresponding to a large number of different samples in one beat being aligned with a single sample in the other. This leads to the following recurrence relation (also shown graphically in Figure 2.4):

$$\gamma(i,j) = d(q_i, c_j) + \min \begin{cases} \gamma(i-1, j-1) \\ d(q_{i-1}, c_j) + \gamma(i-2, j-1) \\ d(q_{i-1}, c_j) + d(q_{i-2}, c_j) + \gamma(i-3, j-1) \\ d(q_i, c_{j-1}) + \gamma(i-1, j-2) \\ d(q_i, c_{j-1}) + d(q_i, c_{j-2} + \gamma(i-1, j-3) \end{cases}$$



(a) Original DTW (b) Constrained DTW

Figure 2.4: Illustration of possible path alignments.

The process described here transforms the original ECG time-series into a sequence of time-warped morphology differences between consecutive beats. To characterize pathological structure within this sequence, we study its spectral characteristics. Since the activity of the heart is quasi-periodic (i.e., since the heart does not beat at an exact rate), the time gap between the samples of the sequence constructed through DTW is not uniform. We address this issue by estimating the power spectral density of the morphology differences time-series using the Lomb-Scargle periodogram [64]. For a time series where the value $m[n]$ is sampled at time $t[n]$, the Lomb-Scargle periodogram estimates the energy at frequency ω as:

$$P(\omega) = \frac{1}{2\sigma^2}\Big(\frac{\sum_n [(m[n]-\mu)\cos\omega(t[n]-\tau)]^2}{\sum_n \cos^2\omega(t[n]-\tau)} + \frac{\sum_n [(m[n]-\mu)\sin\omega(t[n]-\tau)]^2}{\sum_n \sin^2\omega(t[n]-\tau)}\Big)$$

where μ and σ are the mean and variance of the $m[n]$, and τ is defined as :

$$\tan(2\omega\tau) = \frac{\sum_n \sin(2\omega t[n])}{\sum_n \cos(2\omega t[n])}$$

We define our computationally generated biomarker, MV, as energy between 0.30 and 0.55 Hz (as estimated from the Lomb-Scargle periodogram) in the time-series of aggregate morphology changes constructed using DTW. The range of 0.30 to 0.55 Hz is based on theoretical and empirical observations suggesting that the discriminative ability of MV for predicting death following heart attacks is maximized over this range [104]. A flow chart of the whole process for generating MV is shown in Figure 2.5.

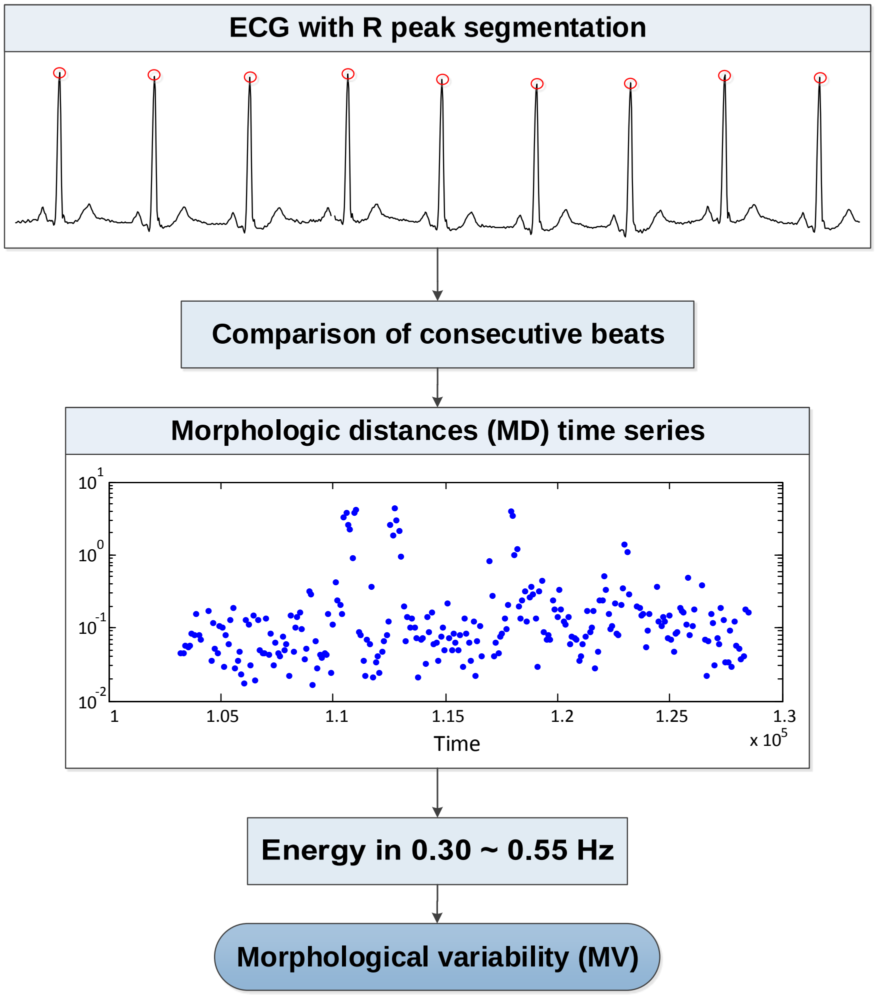

ECG with R peak segmentation
Comparison of consecutive beats
Morphologic distances (MD) time series
10^1
10^0
10^{-1}
10^{-2}
1
1.05
1.1
1.15
1.2
1.25
1.3
Time
x 10^5
Energy in 0.30 ~ 0.55 Hz
Morphological variability (MV)

Figure 2.5: Flow chart of the process for generating Morphological Variability (MV).

CHAPTER III

Improving Existing ECG Biomarkers

3.1 Introduction

In this chapter, we focus on the goal of improving the scalability of existing computational biomarkers from ECG time-series to identify patients at an increased risk of death following ACS. Specifically, MV [104] as described in Section 2.4, has shown to have value in extracting subtle but useful information about the health of the heart from ECG signals. Unfortunately, the existing approach to measure MV is computationally intensive. We aim to address this shortcoming by reducing the computational complexity of MV to scale its use to low-power embedded devices (e.g., ICDs) while maintaining clinically useful discrimination.

The process of measuring MV consists of a modified DTW-based algorithm to quantify time-warped shape deformations in ECG time-series over long periods of time, and a Lomb-Scargle periodogram approach to analyze the resulting non-uniformly sampled time-series representation of aggregate noise in the ECG for pathological structure. To achieve the goal of scaling this basic approach to large databases of long-term ECG time-series, we investigate a novel approach that reduces the quadratic complexity of DTW through an adaptive downsampling of time-series inputs. The use of adaptive downsampling significantly reduces the ECG data presented to DTW while preserving rapidly changing waves (e.g., the QRS-complex) smoothed out by existing downsampling approaches. However, due to adaptively downsampling rapidly changing parts of the ECG time-series less than more slowly changing parts of the signals, this approach also requires changes to the dynamic programming problem underlying DTW. In this chapter, we present solutions for both the goal of adaptively downsampling ECG time-series, and for modifying the DTW dynamic programming formulation to leverage adaptively downsampled inputs.

We evaluate our ideas on data from 765 patients in the DISPERSE2-TIMI33 trial as described in Appendix A.1. Baseline results show that high MV is associated with a 4- to 5-fold increased risk of death within 90 days of a heart attack. Moreover, the use of our proposed adaptive downsampling with a modified DTW formulation achieves an almost 4-fold reduction in runtime relative to DTW, without a significant change in biomarker discrimination. In contrast, existing downsampling approaches obtain a similar reduction in runtime but with noticeably worse performance for risk prediction.

In what follows, section 3.2 reviews previous related work and proposes the concept of adaptive downsampling. Then, section 3.3 details how adaptive downsampling is implemented and how it can be incorporated within the measurement of MV to scale it up to large amounts of long-term time-series data. Section 3.4 presents the evaluation methodology for our study. Section 3.5 discusses the results of our experiments. Section 3.6 offers a summary and conclusions.

3.2 Overview and Previous Work

In order to explore how MV can be scaled for use with very large volumes of ECG data, we first note that the runtime of measuring MV is dominated by the time taken to quantify time-warped morphology differences between consecutive beats. For a total of p beats in an ECG time-series of length less than n, the computational complexity of this step is $O(pn^2)$. While reducing the number of consecutive pairs of beats to be examined (i.e., reducing p) offers one approach to reduce the overall runtime of MV, this approach is made challenging by multiple factors (e.g., poorer estimation of spectral energy, less data available to distinguish between persistent pathological variations and true noise, increased latency for real-time decision-making etc.). As a result, our efforts largely center on addressing the quadratic runtime of DTW (i.e., reducing the n^2 term above).

The basic DTW algorithm quantifies the similarity between two time series with warping by setting up an alignment or correspondence between matching parts of the signals. This process is quadratic in both runtime and space, and does not scale to long time series or large volumes of data. There is a significant body of work focusing on addressing this limitation. Most of the previous work in this area has approached the goal of increasing the efficiency of DTW either by introducing constraints [87], or through dimensionality reduction methods such as the discrete fourier transform [1], singular value decomposition [55], and downsampling by a constant factor [49, 50].

In these existing dimensionality reduction methods, the most popular approach is piecewise aggregate approximation (PAA) [51], which downsamples the original time series by constant factor. An alternate approach is FastDTW [88]. This is a multi-level approach that first finds an optimal warping path in a lower-resolution setting. FastDTW then iteratively expands and improves the solution. Since previous work [26] reports that PAA is the best existing solution to the problem of reducing computational complexity of downsampling, we will focus our evaluation on comparing our approach to PAA.

We note that while the use of downsampling in previous literature improves the runtime and space efficiency of DTW, the decision to carry out this downsampling by a constant factor over time causes both rapidly and slowly changing parts of a signal to be treated similarly. Uniform downsampling using both PAA or FastDTW may be associated with the loss of important information (e.g., sharp R peaks), and specifically hurts our application since MV aims to discover potentially low amplitude disease signatures in the presence of high amplitude baseline activity.

We address this issue by using variable (i.e. adaptive) downsampling. We believe that the process of comparing time-warped signals can be improved by exploiting slowly changing parts of a signal by downsampling them at a higher rate than rapidly changing regions. In contrast to PAA and FastDTW, we therefore propose the idea of adaptive downsampling where the rate of reduction of time series varies according to the rate of changes taking place locally. This allows for the reduction of the number of samples in time series, while retaining sharp changes that would otherwise be smeared if downsampling were applied uniformly to the entire signal. In other words, reduction in the size of the data can be accomplished while preserving information that would otherwise be lost when downsampling by a constant factor.

3.3 Methods

We describe here how to adaptively downsample ECG time-series. This is done by using a trace segmentation-based approach that reduces time-series more in regions where the signals are slowly changing and less in parts of the signal associated with rapid changes. Following this, we describe how to modify the DTW dynamic programming formulation to leverage adaptively downsampled inputs.

3.3.1 Adaptive Downsampling (ADAP)

We achieve adaptive downsampling using trace segmentation [57]. While we describe this approach in more detail subsequently, the basic idea underlying trace segmentation is to divide the signal into regions with equal cumulative derivative activity. This places a higher number of boundaries for downsampling in regions that are rapidly changing (i.e., have higher cumulative derivative activity).

More formally, given a signal $Q = q_1, \ldots, q_n$ and a number of frames θ to downsample this signal to, we first calculate the cumulative difference $D_Q[k]$ for $k = 2, \ldots, n$ between each neighboring pair of samples:

$$D_Q[k] = \sum_{i=2}^{k} \mid q_i - q_{i-1} \mid \tag{3.1}$$

with $D_Q[1] = 0$. The sum of the total differences in Q is given by $D_Q[n]$. The cumulative difference in each adaptive downsampling bin is then set to $d_Q = \frac{D_Q[n]}{\theta}$. Using this, downsampling proceeds by finding the sample numbers t_i for $i = 0, \ldots, \theta$ such that for all values of i we have:

$$t_i = \min\{k \mid D_Q[k] \geq d_Q \cdot i\} \tag{3.2}$$

The corresponding amplitudes of Q at samples t_i are given by $x_i = q_{t_i}$. We can then use interpolation to approximate the fractional sample numbers $\hat{t}_i$ where we would expect $D_Q[\hat{t}_i] = d_Q \cdot i$. For $i = 0, \ldots, \theta$ using the notation:

$$\beta_i = \frac{D_Q[t_i] - d_Q \cdot i}{D_Q[t_i] - D_Q[t_i - 1]} \tag{3.3}$$

we have:

$$\begin{aligned} \hat{t}_i &= t_i - \beta_i \\ \hat{x}_i &= q_{t_i} - \beta_i(q_{t_i} - q_{t_i-1}) \end{aligned} \tag{3.4}$$

The resulting adaptively downsampled representation of the original signal Q is given by two series corresponding to time and amplitude:

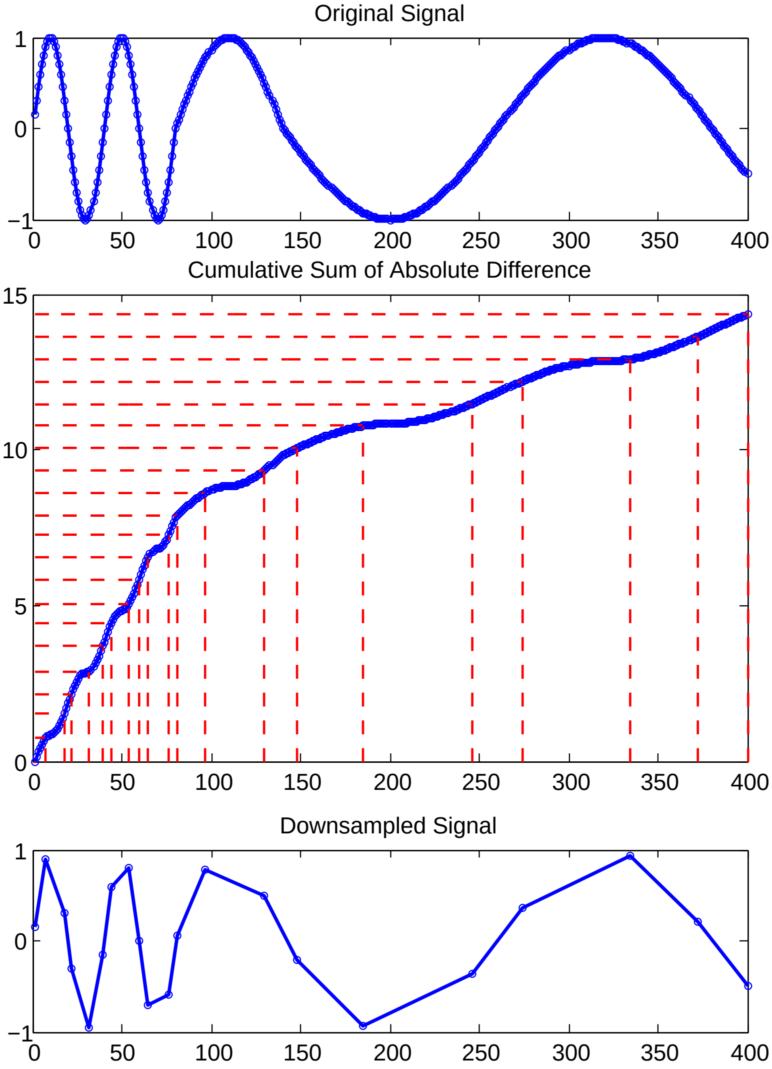
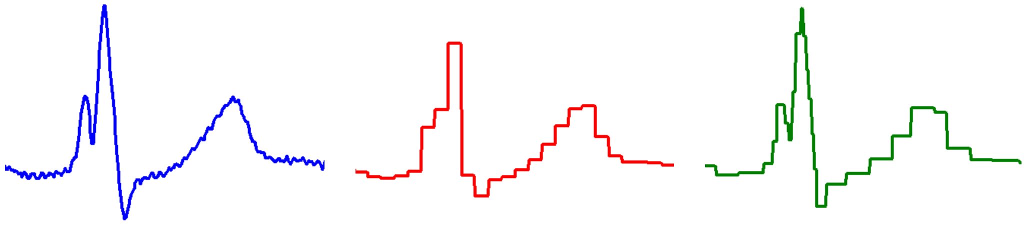


Figure 3.1: An illustration of the trace segmentation process.

$$\begin{aligned} T^Q &= \hat{t}_0, \hat{t}_1, \ldots, \hat{t}_i, \ldots, \hat{t}_\theta \\ X^Q &= \hat{x}_0, \hat{x}_1, \ldots, \hat{x}_i, \ldots, \hat{x}_\theta \end{aligned} \tag{3.5}$$

This process can be carried out in time that is linear in the size of the input. Figure 3.1 presents the trace segmentation approach for downsampling graphically.

For ECG time-series, trace segmentation can preserve important information related to sharply changing parts of the signal (e.g., the QRS-complex). This is illustrated in Figure 3.2. In contrast to PAA, a similar number of adaptively downsampled segments provide a better characterization of notching within the R wave and also the sharpness of the S wave. While PAA achieves good results in a variety of real-world application domains, we believe the distinctions retained by adaptively downsampling are relevant to the specific goal here of measuring MV to predict death following heart attacks.



(a) Original (b) PAA (c) ADAP

Figure 3.2: Adaptive downsampling of ECG signals.

3.3.2 DTW with Adaptive Downsampling

DTW searches for the optimal alignment between two sequences in an efficient manner using dynamic programing. For uniformly downsampled signals, the dynamic programming process is essentially unchanged, although it is applied to reduced representations of the original signals. For adaptively downsampled signals, however, the cost of alignment cannot be calculated in a similarly simple manner from the Euclidean distance between the samples of the downsampled representations. Since the original signal is now divided into segments of variable lengths, this length information needs to be factored into consideration when deriving the distances for the DTW dynamic programming recurrence.

We represent two adaptively downsampled signals Q and C as comprised of segments $s_q(1)$,. . .,$s_q(\theta)$ and $s_c(1)$,. . .,$s_c(\theta)$ respectively, with θ corresponding to the number of downsampled segments. The amplitude of each segment $s_q(i)$ is represented by $x_q(i)$ and the duration by $l_q(i)$ (similar notation is used for the amplitude and duration

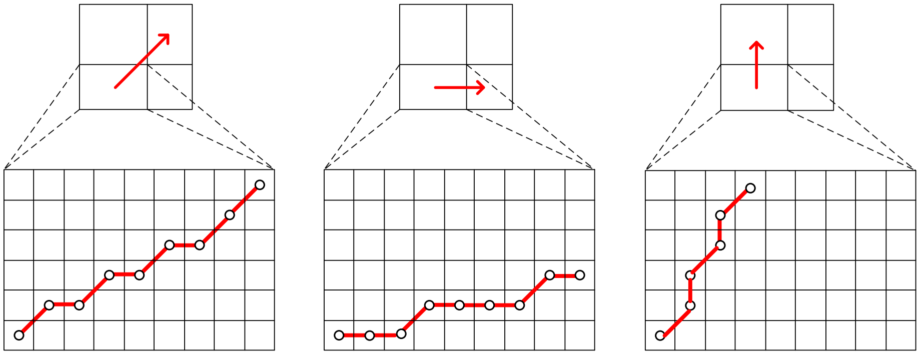


Figure 3.3: Illustration of derivation for adaptive recurrence equation. Left subfigure shows the diagonal case, the middle shows the horizontal case, while the right subfigure shows the vertical case.

of each segment $s_c(i)$). Using this notation, we describe the process through which the dynamic programming of DTW can be modified to handle adaptively downsampled segments.

Figure 3.3 shows, from left to right, three separate possibilities when aligning adaptively downsampled segments. In each case, the alignments of the adaptively downsampled segments are illustrated at the top, and the alignments of the original signals are illustrated below. The leftmost subfigure shows the situation where the adaptively downsampled segments are diagonally aligned, i.e. segment $s_q(i)$ is aligned with segment $s_c(j)$, while $s_q(i+1)$ is aligned with segment $s_c(j+1)$. Intuitively, we expect the warping path between the samples comprising the segments $s_q(i)$ and $s_c(j)$ in the original signals to be close to the diagonal. Without solving for the optimal path of alignment between these original samples, we approximate the cost of alignment between the segments $s_q(i)$ and $s_c(j)$ as the product of $d(s_q(i), s_c(j))$ (i.e., the Euclidean distance of $x_q(i)$ and $x_c(j)$) and $\max(l_q(i), l_c(j))$ (i.e., an estimate for the length of a diagonal path). We adopt a similar approach for the subfigure shown in the middle of Figure 3.3. In this case, the adaptively downsampled segment $s_q(i)$ is aligned with both $s_c(j)$ and $s_c(j+1)$. Again, without solving for the optimal path of alignment between the original samples for these segments, we expect the path of alignment for the samples comprising $s_q(i)$ and $s_c(i)$ to be roughly horizontal. We therefore approximate the length of this path to be the product of $d(s_q(i), s_c(j))$ and $l_q(i)$. The situation shown in the rightmost subfigure (i.e., a roughly vertical path of alignment for the samples comprising $s_q(i)$ and $s_c(j)$) is treated analogously.

We note that our approach of modifying the dynamic programming of DTW for use with adaptive downsampling approximates the path length in each case, and this approximation may not be optimal. However, this approach provides a simple way to augment the dynamic programming of DTW. In particular, in this setting, the recurrence relation for the cumulative path distance $\gamma(i,j)$ till the adaptively downsampled segments i and j can be represented as:

$$\gamma(i,j) = \min \begin{cases} \gamma(i,j,\mathbf{d}) \\ \gamma(i,j,\mathbf{h}) \\ \gamma(i,j,\mathbf{v}) \end{cases}$$

where the cumulative path distance $\gamma(i,j)$ depends on the direction in which the path proceeds next (i.e., diagonal $\mathbf{d}$, horizontal $\mathbf{h}$, or vertical $\mathbf{v}$) and:

$$\gamma(i,j,\mathbf{d}) = d(s_q(i), s_c(j)) \max(l_q(i), l_c(j)) + \min \begin{cases} \gamma(i-1,j-1,\mathbf{d}) \\ \gamma(i-1,j,\mathbf{h}) \\ \gamma(i,j-1,\mathbf{v}) \end{cases}$$

$$\gamma(i,j,\mathbf{h}) = d(s_q(i), s_c(j)) l_q(i) + \min \begin{cases} \gamma(i-1,j-1,\mathbf{d}) \\ \gamma(i-1,j,\mathbf{h}) \\ \gamma(i,j-1,\mathbf{v}) \end{cases}$$

$$\gamma(i,j,\mathbf{v}) = d(s_q(i), s_c(j))l_c(j) + \min \begin{cases} \gamma(i-1,j-1,\mathbf{d}) \\ \gamma(i-1,j,\mathbf{h}) \\ \gamma(i,j-1,\mathbf{v}) \end{cases}$$

Path Constraints: Recall that in Section 2.4 we described a modification to the basic DTW recurrence relation to find more biologically plausible alignments (i.e., the situation shown in Figure 2.4). In this case, we made use of the recurrence relation:

$$\gamma(i,j) = d(q_i, c_j) + \min \begin{cases} \gamma(i-1,j-1) \\ d(q_{i-1}, c_j) + \gamma(i-2,j-1) \\ d(q_{i-1}, c_j) + d(q_{i-2}, c_j) + \gamma(i-3,j-1) \\ d(q_i, c_{j-1}) + \gamma(i-1,j-2) \\ d(q_i, c_{j-1}) + d(q_i, c_{j-2} + \gamma(i-1,j-3) \end{cases}$$

We adopt an analogous approach to constrain DTW with adaptive downsampling for more meaningful alignments. Since the original signal is divided into unequally sized segments, we note that the above recurrence would not be directly applicable. Instead of restricting valid paths to pass through no more than 3 consecutive horizontal or vertical steps, we therefore restrict the path to traverse through at most k steps such that no such implausible alignment would occur. In other words, a segment $s_q(i)$ is only allowed to align with segments of s_c such that the total length of those k segments is no greater than three times the length of $s_q(i)$, which can be expressed as $3 \cdot l_q(i) \leq \sum_{n=1}^{k} l_c(j-n)$.

3.4 Evaluation

We evaluated our research on ECG data from patients in the DISPERSE2-TIMI33 trial. In our study, we used data from the first 24 hours of ECG recording during hospitalization to predict the risk of death following heart attacks. There were a total of 765 patients in the DISPERSE2-TIMI33 trial with available ECG signals sampled at 128Hz used in this analyses, with 14 deaths during 90-day follow-up.

We compared the basic MV algorithm to MV measured with downsampling using PAA and to MV measured with adaptive downsampling. This comparison was performed in multiple ways. First, we measured the areas under the receiver operating characteristic curves (AUROCs, see detailed discussion in Appendix B.2) for all three approaches. As part of this evaluation, we compared the AUROC values for the downsampled MV approaches to the basic MV algorithm without downsampling using the method proposed by DeLong et al. [29] to assess whether the changes are statistically significant. Second, we also assessed the MV models with downsampling relative to the basic MV algorithm without downsampling by measuring the integrated discrimination improvement (IDI, refer to Appendix B.3 for a detailed description) proposed by Pencina et al. [83]. This was done by translating the MV values obtained through each approach into regression-based probabilistic risk estimates, and then measuring the difference between the mean predicted probabilities of events and non-events.

In addition to evaluating the predictive accuracy of MV measured through each approach, we also evaluated the runtime of the algorithms as the average time taken across ten runs to compute MV for all patients. These experiments were performed on a machine with quad-core Intel Xeon X3450 processors (2.67 GHz, 8MB Cache) and 8 GB RAM. The distance metrics were uniformly implemented in C++ on the Red Hat Enterprise Linux Server release 5.6 (Tikanga).

Finally, we also assessed how the relative ranking of patients between the different MV approaches changed with adaptive and non-adaptive downsampling. This

Table 3.1: Univariate association of MV and other clinical variables with death following heart attacks.

Parameter	Hazard Ratio	P-value
Age>65	3.72	0.024
Women	2.76	0.054
Smoker	0.53	0.225
Hypertension	6.66	0.067
Diabetes	2.77	0.049
Hyperlipidemia	0.66	0.422
Previous Heart Attack	1.94	0.210
Previous Angina	2.86	0.103
ST depression>0.5mm	2.69	0.091
MV	5.16	0.002

metric was used to study how downsampling moves patients relative to each other while measuring MV. To measure this information, we computed the average absolute difference in the ranking of each patient by MV across different approaches.

3.5 Results

3.5.1 Clinical Utility

The basic MV algorithm achieved an AUROC of 0.75 for discriminating between high and low risk patients following heart attacks. When the MV predictions were dichotomized at a simple threshold (MV>50 vs. MV≤50), patients with high MV were found to be at a significantly increased risk of death following heart attacks (Figure 3.4). For comparison, we show the relative increase in risk between patients with high and low MV, as well as the relative increases in risk for a variety of existing clinical variables in Table 3.1. We use the hazard ratio as a measure of relative increase in risk. In the DISPERSE2-TIMI33 dataset, MV identified a group of patients at a higher relative risk than any of these other metrics. These results are consistent with the earlier findings about MV reported in the clinical literature [104].

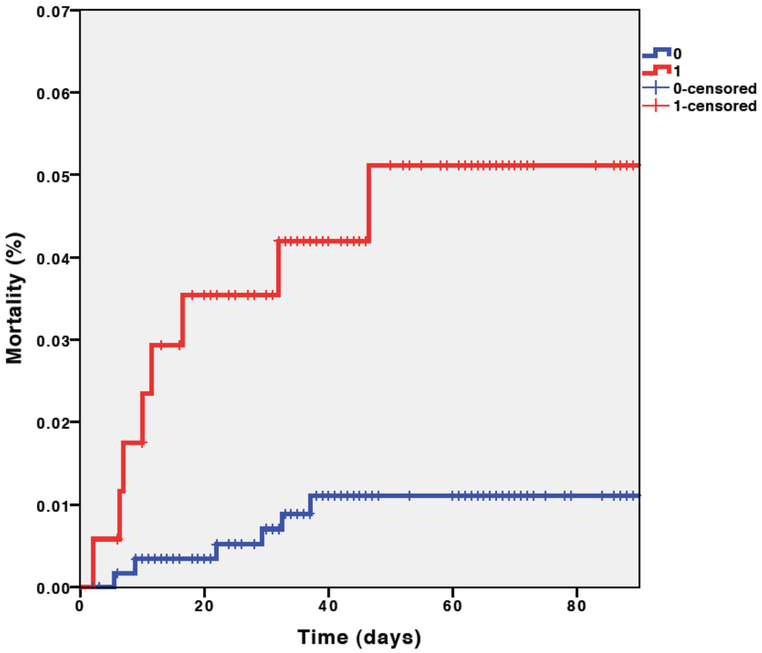


Figure 3.4: Kaplan-Meier mortality curve for patients in high MV (MV>50; shown in red) and low MV (MV≤50; shown in blue) groups. Patients with high MV were at a consistently elevated risk of death over the 90 day period following a heart attack.

3.5.2 Computational Efficiency

Table 3.2 compares the AUROC for the basic MV algorithm with the AUROCs obtained for MV measured with downsampling using PAA and MV measured with adaptive downsampling. For both the downsampling approaches, we experimented with downsampling the original heart beat signals to down to 30, 50 and 70 samples. In order to further evaluate if the differences in the AUROC estimates for different models has significance, we applied Delong's method [29], which is an asymptotically exact method to evaluate the uncertainty of an AUROC and of comparisons between two AUROCs.

In general, downsampling the original signal led to a reduction in the discriminative ability of MV (although this difference was not significant at the 5% level given the sample size). In all of our experiments, however, MV with adaptive downsampling achieved a higher AUROC than downsampling with PAA for a similar factor

Table 3.2: Comparison of AUROCs between DTW, PAA-DTW, and ADAP-DTW

Methods	AUROC	P-value
DTW	0.748	Referent
PAA_{30}	0.658	0.331
PAA_{50}	0.669	0.345
PAA_{70}	0.693	0.384
$ADAP_{30}$	0.718	0.345
$ADAP_{50}$	0.737	0.729
$ADAP_{70}$	0.736	0.721

Table 3.3: IDI comparing DTW with PAA-DTW and ADAP-DTW

Methods	IDI	P-value
PAA_{30}	-0.009	0.164
PAA_{50}	-0.009	0.178
PAA_{70}	-0.007	0.209
$ADAP_{30}$	0.017	0.160
$ADAP_{50}$	0.014	0.204
$ADAP_{70}$	0.020	0.160

of reduction. These results suggest that our use of adaptive downsampling retained more information that was relevant to the task of distinguishing between high and low risk patients than the use of PAA for this application.

We also studied changes in the clinical utility of MV with downsampling (as assessed by the IDI) for each downsampled approaches relative to the basic MV algorithm. These results are presented in Table 3.3. In this case, the data from our experiments show that (consistent with the AUROC case) the use of downsampling with PAA led to a small decrease in performance. Conversely, the use of adaptive downsampling actually resulted in an increase in discriminative performance relative to the basic DTW algorithm as measured by the IDI. The differences for both downsampling with PAA and with adaptive downsampling relative to the basic DTW algorithm were not significant at the 5% level given the sample size.

The relative changes in ranks of patients between the basic DTW algorithm and the DTW approaches with downsampling are shown in Table 3.4. Consistent with the AUROC and IDI results, DTW with adaptive downsampling resulted in a smaller

Table 3.4: Average change in patient ranks relative to the basic DTW algorithm (shown as percentages of the DISPERSE2-TIMI33 population).

Methods	Change
PAA_{30}	20.3%
PAA_{50}	21.2%
PAA_{70}	20.9%
ADAP_{30}	18.3%
ADAP_{50}	18.8%
ADAP_{70}	18.4%

Table 3.5: Computation time of the different MV algorithms.

Methods	Time (sec)
DTW	146,940
PAA_{30}	5,663
PAA_{50}	7,311
PAA_{70}	9,207
ADAP_{30}	7,831
ADAP_{50}	13,292
ADAP_{70}	20,782

relative change in rank within the DISPERSE2-TIMI33 population relative to the basic DTW algorithm.

These empirical improvements are supported by a theoretical analysis of the runtime of DTW with adaptive downsampling. In the case of the original DTW without window constraints, our approach improves performance to $O((\frac{n}{k})^2)$ from $O(n^2)$. In the case where DTW is constrained to use a window of size w, average performance is improved to $O(nw/k)$, while worst case performance is $O(nw)$. We find that performance improves in practice as seen for the timing results for the different methods in Table 3.5. While downsampling reduced the runtime of the basic DTW algorithm substantially in each case, this reduction was greater for PAA than with the use of adaptive downsampling. This result can be attributed to the additional work that needs to be done to solve the modified dynamic programming problem for adaptively downsampled DTW. Comparing the PAA and adaptively downsampled approaches based on time rather than downsampling factor, however, still showed a higher level of performance with adaptive downsampling than with the use of PAA (e.g., for PAA_{70} AUROC: 0.693, IDI: -0.007, average rank change: 20.9% and time: 9,207 vs. for ADAP_{30} AUROC: 0.718, IDI: 0.017, average rank change: 18.3% and time: 7,831)

3.6 Conclusion

In this chapter, we explored mining the noise-like variations in long-term ECG time-series to identify patients at an increased risk of death following heart attacks. To achieve this, we described a modified DTW- and Lomb-Scargle periodogram-based approach that first transforms ECG time-series into sequences of beat-to-beat time-aligned morphology differences, and then relates properties of these sequences to patient risk. While the ideas underlying this work derive from earlier experiments reported in the clinical literature [104], we focused here on the question of how this basic approach can be scaled to very large ECG time-series databases. As part of this work, we investigated a novel approach to address the quadratic runtime of DTW. In particular, we proposed the idea of adaptive downsampling, i.e., downsampling slowly changing parts of a signal much more than rapidly changing parts of the same signal, to reduce the size of the inputs presented to DTW while retaining a good representation of the original time-series being compared. We also described changes to the dynamic programming underlying DTW to exploit such adaptively downsampled signals, where the downsampled segments may be of varying lengths.

We evaluated our ideas on real-world data from patients within the DISPERSE2-TIMI33 trial. Our experiments suggest that measuring MV with adaptive downsampling substantially reduces runtime while providing similar performance to the basic MV algorithm that is not optimized for large volumes of data. In addition, the use of adaptive downsampling leads to more accurate performance than downsampling through the commonly used approach of PAA. Finally, we note that while the discussion here focuses primarily on long-term ECG, we have also applied the concept of adaptive downsampling on broader set of data. The results of this investigation show consistent improvements in runtime when using adaptive downsampling; suggesting the general applicability of this approach to time series [26].

We conclude this section by observing that further exploration is needed to determine whether modifications of our approach can yield significantly better results. As one example, it may be possible to apply weights to adaptive downsampled segments based on their length so that longer downsampled segments are assigned lower importance. Similarly, it may also be possible to jointly optimize the adaptive downsampling-based DTW approach by combining information in the time-domain with information in the frequency-domain to obtain better performance. These ideas need to be explored in more detail subsequently to fully characterize the opportunities to improve the measurement of MV.

CHAPTER IV

Developing New ECG Biomarkers

4.1 Introduction

Building upon the work in Chapter III where we improve on existing biomarkers, in this chapter, we aim to develop novel computational biomarkers to risk stratify patients for death following coronary attacks. Specifically, we focus on using electrocardiographic (ECG) data from a large patient population to discover heart rate patterns that are statistically overrepresented or underrepresented in patients who died in the months immediately following a heart attack or unstable angina relative to patients who survive this period. We propose a randomized hashing- and greedy centroid selection-based algorithm to efficiently discover such heart rate patterns in large high-resolution ECG datasets captured continuously over long periods from thousands of patients. The discovery of those patterns can further help us in the bigger goal of improving clinical cost-benefit analyses to determine therapies most appropriate for individual cases.

We evaluate our work on data from over 3,000 patients from two separate cohorts of patients admitted to the hospital following coronary attacks, and show that our computationally generated biomarkers can correctly identify patients at high risk of death, even after adjusting for information in existing risk stratification metrics. We note that while our investigation is focused on the specific clinical application of cardiovascular risk stratification, the techniques we propose can be applied more broadly to other problems related to approximate sequential pattern discovery in large datasets.

The remainder of this chapter is organized as follows. Section 4.2 first describes how previous work focusing on heart rate time series is inadequate. Sections 4.3 and 4.4 then describe how heart rate time series can be abstracted into symbolic sequences and how problems relevant to clinical stratification can be framed and solved using this abstraction. Section 4.5 details the evaluation methodology for this work, and the results of this investigation are presented in Section 4.6. We conclude with a summary and discussion of these results in Section 4.7.

4.2 Overview and Previous Work

There is an extensive body of research focused on the analysis of heart rate time-series over long periods [79], especially, time- and frequency-domain measures of HRV. However, most of the existing works to HRV only looks at aggregated variability instead of trying to extract short-term structure in heart rate. Here, we advance these efforts by identifying specific patterns (i.e. short-term structures) of heart rate changes that may be used for risk stratification. These patterns may correspond to activity that is either overrepresented in the patients who experience adverse outcomes (i.e., patterns associated with the causal disease mechanisms) or overrepresented in patients who remained event free (i.e., patterns associated with protective mechanisms). These patterns can then be used to develop risk stratification models that score patients along a risk continuum.

Our search for these specific patterns supplements traditional HRV analyses [67] in two ways: (1) our research diverges from the typical approach of quantifying aggregate variability through simple time- and frequency-domain metrics (Table 2.1) by extracting more specific patterns associated with elevated cardiovascular risk (i.e., our research identifies structure within aggregate variability measured by existing metrics), and (2) our research provides a more complete assessment of information in heart rate by capturing patterns associated with both low and high variability. While most HRV research focuses narrowly on the low variability case, our proposed algorithms can also identify patterns associated with high risk increased variability. There is recent evidence suggesting that such a broader focus can provide a more complete assessment of cardiovascular health [103, 106].

Despite the potential clinical utility of ECG-based heart rate markers, discovering these patterns is difficult due to three factors. First, the sheer volume of available data poses a serious challenge. For example, the ECG signals from just a single patient admitted to a hospital following a coronary attack would fill up thousands of pages. Our research attempts to find patterns in ECG data collected continuously from thousands of such patients over days to weeks following a coronary attack. This creates space and runtime challenges at both the algorithmic and platform level. Second, due to the presence of noise, our research explores patterns that are approximate, i.e., where the same heart rate sequence can occur in many parts of the same signal or across different patients in an imperfectly conversed form. Third, there is considerable variation between patients, and physiological information must be registered across individuals during the pattern discovery process.

Our research addresses these challenges by re-expressing heart rate time-series in a symbolic form where the symbols have consistent meaning across patients. We then approach the goal of discovering approximate heart rate patterns within these symbolic sequences by proposing a new motif discovery algorithm that makes use of on-line greedy centroid selection and randomized hashing to identify statistically interesting sequential activity in positively and negatively labeled examples. We combine information from these patterns in survival models that can be used for risk stratification.

4.3 Creating Symbolic Heart Rate Sequences

Given the ECG signals $X_i[n]$ for patients $i = 1, \ldots, N$, we start by first extracting the heart rate from these time series. To segment the ECG signals into beats, we use two open-source QRS detection [41, 115]. QRS-complexes are marked at locations where both algorithms agree. The time interval between the QRS-complexes measures the length of each heart beat, and can consequently be used to measure the instantaneous heart rate in beats per minute.

Denoting the instantaneous heart rate time series as $Z_i[n]$ for $i = 1, \ldots, N$, we then symbolize the heart rate for each patient using symbolic aggregate approximation (SAX) [62]. Basically, every one heart beat gets symbolized into one symbol. This is done by first partitioning the heart rate measurements within each patient's time series into equiprobable bins, and then assigning each heart rate measurement with a symbol corresponding to the index number of its bin (where the index number '1' corresponds to the equiprobable bin with the lowest mean heart rate measurements, '2' corresponds to the equiprobable bin with the next lowest mean heart rate measurements and so on). The number of bins determines the size of the symbol alphabet. In our work, we choose an alphabet size of 4 for SAX, corresponding to an abstraction of the heart rate into low, moderately low, moderately high, and high categories. Under this symbolization, the distance between symbol '1' vs. '2' and distance between symbol '1' vs. '3' is different.

We note that while clinical definitions (e.g., bradycardia corresponding to heart rate below 60 beats per minute, tachycardia corresponding to heart rate above 100 beats per minute, and normal heart rate between 60-100 beats per minute) can also be used to achieve a discretization of the heart rate time series for each patient, our use of SAX has the advantage of providing a layer of normalization across patients. Despite baseline differences in heart rate between patients, the symbols derived through SAX have consistent meaning across the population. We represent the resulting symbolic heart rate sequences as $S_i[n]$ for $i = 1, \ldots, N$.

4.4 Pattern Discovery in Symbolic Sequences

We now discuss related literature on pattern discovery and then formulate our problem of discovering over-represented patterns in symbolic sequences and describe an algorithm to efficiently discover such patterns.

4.4.1 Existing Methods

Previous work in computational biology tries to find conserved patterns that are unlikely to occur by chance but are encountered repeatedly. This is done using approaches including two component mixture (TCM) [10], Gibbs sampling [108], and Consensus [99]. In these efforts, there is no focus on discrimination. In the data mining community there are also other approaches (e.g., shapelets [112]) that are focused on finding primitive representative patterns. The focus in these cases again is on finding patterns that occur more frequently than would be expected purely by change, that is, over-common patterns within a single population (as opposed to patterns that occur differently between two separate populations).

In short, existing approaches do not use both positive and negative examples to find motifs that occur with differential distribution across patients with and without outcomes, which is understandable given that existing approaches in computational biology attempt to find motifs that are likely to have functional significance within the genome; an analysis which is not comparative in nature. In ou work, we explore a different problem formulation from these efforts.

4.4.2 Problem Formulation

We frame the problem of discovering heart rate patterns that have value in cardiovascular risk stratification as:

PROBLEM FORMULATION 1. *Given two sets of sequences* $S^+ = \{S_i^+ | i = 1, \ldots, N^+\}$ *and* $S^- = \{S_i^- | i = 1, \ldots, N^-\}$ *drawn from families* F^+ *and* F^-*, such that* $F^+ \cap F^- = \oslash$*, find all subsequences of length* L *that occur in an approximate form with high relative likelihood in either* F^+ *or* F^-*.*

where the approximate form of a subsequence $A = a_1, \ldots, a_L$ corresponds to the subsequence A and all other subsequences $B = b_1, \ldots, b_L$ such that:

$$\max_{i=1,\ldots,L-W+1} [l_p(b_i, \ldots, b_{i+W-1}; a_i, \ldots, a_{i+W-1})] \leq d$$

and where the l_p distance is defined as:

$$l_p(b_i, \ldots, b_{i+W-1}; a_i, \ldots, a_{i+W-1}) = (\sum_{j=i}^{i+W-1} |a_j - b_j|^p)^{1/p}$$

This notion of an approximate pattern is more natural than the use of the l_p norm to directly assess the distance between patterns (i.e., an approximate pattern consists of A and all subsequences B such that $l_p(A, B) \leq d$). This is because our notion of an approximate pattern prevents matches from being substantially different at any local point along the subsequences. Moreover, it defines distance as a function of pattern length, and allows for the pattern discovery process to use a single parameter while searching for shorter or longer patterns. We observe, however, that the techniques presented in this paper can be applied just as easily to the case where the l_p norm is used instead of the definition of approximate matches above.

We defer the question of how to find approximate patterns for the moment and start by describing how a similar problem to find exact patterns can be solved. In this setting, a simple hash table-based approach can be used to make a linear pass through all sequences in S^+ and S^- both to identify the unique subsequences U_i for $i = 1, \ldots, M$ present in the entire dataset and to measure the frequency with which each of these unique subsequences occurs in positive and negative examples.

The resulting frequencies for the U_i in each of the sequences in S^+ and S^- can then denoted by the vectors:

$$f_i^+ = \{f_{i,k}^+ | k \in S^+\}$$
$$f_i^- = \{f_{i,k}^- | k \in S^-\}$$

Using these frequencies, i.e., the normalized occurrences of U_i in each sequence in S^+ and S^-, the subsequences that are overrepresented in either positive or negative examples can be found through different approaches. For example, rank sum testing or the AUROC can both be used to identify the U_i that successfully distinguish between positive and negative examples. The AUROC, in particular, is widely used in many different applications and is considered the standard in medicine for evaluating risk stratification methods. The AUROC can be interpreted as the probability that for a pair of randomly chosen comparable examples from S^+ and S^-, U_i occurs more frequently in the positive example than the negative one. Comparable examples correspond to pairs where the frequency of U_i differs between the examples. AUROC values that are high (i.e., close to 1) or low (i.e., close to 0) both reflect subsequences that have high discriminative value.

One limitation of this approach, however, is that it does not capture information related to the timing of labels. In settings such as clinical risk stratification, where patients are only monitored for a specific period, and may occasionally leave a study before it is complete (i.e., the phenomenon of censoring), there is additional information in the labels that this process fails to consider. For example, a patient who leaves a three-month study at the end of the first month would have the label of being event free despite potentially experiencing death before the end of the study period. Situations such as these, where survival characteristics are also important in addition to labels, can be addressed by a slightly revised problem statement:

PROBLEM FORMULATION 2. *Given two sets of sequences $S^+ = \{S_i^+ | i = 1, \ldots, N^+\}$ and $S^- = \{S_i^- | i = 1, \ldots, N^-\}$ drawn from families F^+ and F^-, such that $F^+ \cap F^- = \oslash$ and with event times $T^+ = \{T_i^+ | i = 1, \ldots, N^+\}$ for sequences in S^+ and censoring times $T^- = \{T_i^- | i = 1, \ldots, N^-\}$ for sequences in S^-, find all subsequences of length L that occur in an approximate form with high relative concordance in either F^+ or F^-.*

where the notion of an approximate patterns is similar to the first problem formulation.

In this case, the process of finding exact patterns is identical to the earlier problem formulation except for the concordance index (C-index) being used to identify the U_i that successfully distinguish between positive and negative examples. The C-index is similar to the AUROC, and can be interpreted as the probability that for a given pair of randomly chosen comparable examples, U_i occurs more frequently in the example that experiences an event before the other example. Comparable examples in this case must both be positive samples, or one positive sample paired up with a negative sample such that the positive event occurs before the censoring time for the negative sample.

4.4.3 Discovering Approximate Patterns

Our discussion so far has focused on the discovery of exact patterns for risk stratification. Both problem formulations are similar in the initial step of measuring the frequency with which the subsequences U_i occur exactly in S^+ and S^-, and differ only in their use of the AUROC or C-index. To find approximate patterns, we extend this approach. We first efficiently measure the frequency with which the subsequences U_i occur in an approximate form in S^+ and S^-, and then assess the predictive ability of these approximate patterns using the AUROC or C-index. The goal of the ideas presented in the remainder of this section is to provide a way to efficiently carry out the first step of measuring the frequency of approximate patterns in S^+ and S^-, i.e., to compute:

$$\hat{f}_i^+ = \sum_{j \in D_i} f_j^+$$

$$\hat{f}_i^- = \sum_{j \in D_i} f_j^-$$

where $\hat{f}_i^+$ and $\hat{f}_i^-$ are the vectors obtained by summing up the vectors f_i^+ and f_i^- within the approximate neighborhood D_i of U_i.

One approach to efficiently measure the frequency of approximate patterns in S^+ and S^- is to use a simple hash table-based algorithm to first determine f_i^+ and f_i^- for all U_i, and to then compare each U_i with all other unique patterns to detect its approximate neighborhood D_i and compute $\hat{f}_i^+$ and $\hat{f}_i^-$. The resulting $\hat{f}_i^+$ and $\hat{f}_i^-$ can be used to assess the importance of the approximate pattern centered at U_i through either the AUROC or the C-index. This approach is associated with two limitations. First, matching each of the M unique U_i against all the other $M - 1$ unique subsequences requires $O(M^2)$ time, which is prohibitively expensive for large M. For extremely large datasets, the value of M may be close to Λ^L where Λ is the size of the alphabet used for symbolization. Second, the neighborhoods of approximate patterns may overlap substantially, leading to the best results returned by this process essentially corresponding to the same underlying behavior.

To address these limitations we propose an algorithm to measure the frequency of approximate patterns based on randomized hashing and greedy centroid selection. Each of these ideas is described in more detail subsequently.

Locality Sensitive Hashing with l_p Distance: We note that the goal of matching each U_i with other subsequences that lie within its approximate neighborhood can be reduced to the goal of first identifying a small number of candidate subsequences that may be potential matches, and then pruning away these candidates for actual matches. Being able to efficiently identify the first set of potential matches in this setting can greatly decrease the overall runtime of pattern discovery.

To achieve this, we build upon the observation that given the definition of an approximate match in Section 4.4.2 (i.e., subsequences are approximate matches if they have an l_p distance of at most d over any window of length W), the total l_p distance any potential match can have from a given U_i is bounded by $\gamma = \frac{dL}{W}$. We therefore focus on reducing the runtime of the pattern discovery process by efficiently finding all subsequences within a distance γ of each U_i, and pruning away from this set any subsequences that are not true approximate matches. This goal of finding all subsequences within a l_p radius of γ is related to the R-near neighbor reporting problem, that is, for a query point q_i find all points q_j within a radius R of q_i [4]. In our case, the problem of finding the approximate neighborhood of each U_i corresponds to solving the R-near neighbor reporting problem for each U_i with $R = \gamma$. We achieve this in a computationally efficient manner through locality sensitive hashing (LSH) [4, 28].

We briefly review LSH here for the purpose of completeness. The key idea of LSH is to hash data points using several hash functions with the property that for each function, the probability of collision is much higher for objects that are close to each other than for those that are far apart. This allows for the efficient discovery of nearest neighbors by hashing a query point and searching only through elements stored in buckets containing that point. In addition, since LSH is a hashing-based scheme, it can be naturally extended to dynamic datasets where insertion and deletion operations need to be supported.

More formally, let $\mathcal{H}$ be a family of hash functions mapping $\mathbb{R}^L$ to some universe ζ. For any two points $q_i \in \mathbb{R}^L$ and $q_j \in \mathbb{R}^L$, we can choose a function h from $\mathcal{H}$ uniformly at random and analyze the probability that $h(q_i) = h(q_j)$. The family $\mathcal{H}$

is called locality sensitive [28] if it satisfies the following conditions given a distance measure Θ:

DEFINITION 1. *A function family* $\mathcal{H} = \{h : \mathbb{R}^L \to \zeta\}$ *is called* (R, cR, p_1, p_2)*-sensitive if for any two points* $q_i, q_j \in \mathbb{R}^L$

- *if* $\Theta(q_i, q_j) \leq R$ *then* $Pr_{\mathcal{H}}[h(q_i) = h(q_j)] \geq p_1$
- *if* $\Theta(q_i, q_j) > cR$ *then* $Pr_{\mathcal{H}}[h(q_i) = h(q_j)] \leq p_2$

For an LSH function family to be useful it has to satisfy the inequalities $c > 1$ and $p_1 > p_2$. In this case, the LSH family can be used to efficiently solve the R-near neighbor reporting problem [4]:

DEFINITION 2. *Given a set of* Q *points in an* L*-dimensional space* $\mathbb{R}^L$ *and parameters* $R > 0, \delta > 0$*, construct a data structure that given any query point* q_i*, reports each* R*-near neighbor of* q_i *in* Q *with probability* $1 - \delta$

Typically, one cannot use $\mathcal{H}$ directly, since the gap between p_1 and p_2 may be small. An amplification process is used to achieve a desired probability of collision. This amplification process involves concatenating several functions chosen from $\mathcal{H}$.

The basic LSH indexing method can be described as follows [4]. For an integer k, we first define the function family $\mathcal{G} = \{g : \mathbb{R}^L \to \zeta^k\}$ such that $g \in \mathcal{G}$ is given by $g(q_i) = (h_1(q_i), \ldots, h_k(q_i))$ where $h_j \in \mathcal{H}$ for $1 \leq j \leq k$ (i.e., g is the concatenation of k LSH functions). For an integer ω, we then choose $g_1, \ldots, g_\omega$ from $\mathcal{G}$ independently and uniformly at random. Each of these functions g_i for $1 \leq i \leq \omega$ is used to construct one hash table where all the elements in Q are hashed using g_i. This data structure, comprising ω hash tables in total, is used to find matches to queries. Given a query q_i, the first step is to generate a candidate set of neighbors by the union of all buckets that the query q_i is hashed to. False positives are then removed from this candidate set.

Intuitively, concatenating multiple LSH functions to produce each g_i makes the probability of distant objects colliding small. However, it also reduces the collision probability of nearby objects. This results in the need to create and query multiple hash tables constructed with different g_i.

Different LSH families can be used for different choices of Θ. In our work, we make use of the LSH families based on p-stable distributions for l_p norms [28]:

DEFINITION 3. *A distribution Γ is called p-stable if there exists $p \geq 0$ such that for any n real numbers $v_1, \ldots, v_n$ and i.i.d. variables $Y_1, \ldots, Y_n$ drawn from Γ, the random variable $\sum_i v_i Y_i$ has the same distribution as the variable $(\sum_i |v_i|^p)^{1/p} Y$ where Y is a random variable with distribution Γ.*

p-stable distributions can be used to generate hash functions that obey the locality sensitive property. Given a random vector a of dimension L whose each entry is chosen independently from a p-stable distribution, the dot product of two vectors v_1 and v_2 with a projects these vectors onto the real line. It follows from p-stability that for the vectors, the distance between their projects $(a.v_1 - a.v_2)$ is distributed as $\Theta(v_1, v_2)Y$ where Y is a random variable drawn from a p-stable distribution. If the real line is divided into equi-width segments, and vectors are assigned hash values based on which segments they project onto when taking the dot product with a, then it is clear that this hash function will be locality preserving.

More formally, we can define hash functions based on this idea as [28]:

$$h_{a,b}(v) = \lfloor \frac{a.v + b}{C} \rfloor$$

where a is an L dimensional vector with entries chosen independently from a p-stable distribution as described above, and b is a real number chosen uniformly from the range $[0, C]$. Each hash function $h_{a,b} : \mathbb{R}^L \rightarrow \mathbb{Z}^+$ maps a vector v onto the set of positive integers. For the $p = 2$ case (i.e., the l_2 norm corresponding to the Euclidean distance metric), these hash functions can be created using the Gaussian distribution, which is known to be 2-stable (for the l_1 norm the Cauchy distribution can be used).

For any function g_o, the probability that $g_o(q_i) = g_o(q_j)$ where q_j is an R-neighbor of q_i, is at least p_1^k. The probability that $g_o(q_i) = g_o(q_j)$ for some $o = 1, \ldots, \omega$ is then at least $1 - (1 - p_1^k)^\omega$. If we set $\omega = \lceil \log_{1-p_1^k} \delta \rceil$ so that $(1 - p_1^k)^\omega \leq \delta$, then any R-neighbor of q_i is returned by the algorithm with probability at least $1 - \delta$ [4]. While the worst case performance of this approach is linear, the algorithm is typically sublinear on many datasets.

To choose k, we note that while larger values of k lead to hash functions that are more selective, they also necessitate more hash tables ω to reduce false negatives. To address this tradeoff, between hash functions that lead to smaller hash table buckets but more hash tables, and hash functions that lead to larger hash table buckets but fewer hash tables, we make use of a practical approach that is often recommended [4] to optimize the parameter k. This involves a preliminary training phase using a small number of data points and a set of sample queries. The value of k that provides the best performance is used to develop the LSH data structure.

Greedy Centroid Selection: LSH makes the search for the approximate neighbors of each U_i more efficient but does not address the issue of overlap between patterns. We note that this overlap affects both the quality of the results (i.e., the situation where the top results correspond to slight variations of a single pattern) as well as the runtime of the pattern discovery process (i.e., redundant work being performed to assess very similar patterns repeatedly).

We address this issue by searching for centroids that cover all the subsequences U_i in the dataset and provide a more compact representation of the space of potential patterns. The R-near neighbor reporting problem can then be focused on finding the approximate neighbors of these centroids.

One approach to identify these centroids is to use clustering. However, finding these centroids by clustering all the U_i is prohibitively expensive for large values of M. In fact, we note that the computational challenges of clustering all the U_i are analogous to the computational challenges of solving the R-near neighbor reporting problem for all the U_i in the first place.

We therefore adopt a greedy approach to select centroids. We maintain a working set of centroids (initialized to $\{U_1\}$) using an LSH data structure and perform the following test on each new U_i encountered during a linear scan of the sequences in S^+ and S^- (as described in Section 4.4.2). If the new U_i has any match in the working set of centroids, then the new U_i is ignored. Otherwise, it is added to the working set of centroids. Since the test at each step involves finding *any* match in the working set rather than *all* matches (i.e., the R-near neighbor problem as opposed to the R-nearest neighbor reporting problem), the LSH data structure is extremely efficient for this task [4]. Moreover, the ability of LSH to add points dynamically to the maintained working set also makes it well-suited to the greedy selection of centroids.

It is important to point out that the greedy centroid selection process serves only to identify centroids that should be analyzed more thoroughly using the LSH-based R-near neighbor reporting process described earlier. This is because while the selection process does find matches between centroids and other unique subsequences in the data, these matches represent an incomplete set that must be augmented subsequently.

Optimizations: In addition to the algorithm-level improvements introduced by LSH and greedy centroid selection, we also make use of various platform-level optimizations. For example, in order to fit the data structures used for the pattern discovery process in memory, we choose to retain only non-zero elements within the vectors f_i^+ and f_i^- (and similarly in $\hat{f}_i^+$ and $\hat{f}_i^-$). This is due to the occurrence of subsequences generally being sparse within both S^+ and S^- for large values of L. In addition, we further improve memory usage by retaining pointers to the first incidence of each U_i in the data, rather than maintaining separate copies of each subsequence.

4.4.4 Integrating Patterns into Risk Models

The algorithm described in Section 4.4 can be used to discover heart rate patterns containing useful information for risk stratification, ordered by either AUROC (Formulation 1) or C-index (Formulation 2). We combine the information in these patterns to create composite models that can be applied for risk assessment. Specifically, we select the top Γ uncorrelated patterns in a greedy manner, ordered by AUROC or C-index, and combine them in either a logistic regression or Cox proportional hazards regression model to predict risk. The models can be further enriched by combining patterns of multiple lengths. While a number of other methods can be used for the goal of combining patterns, including algorithms recently proposed in the machine learning and data mining literature, our choice of logistic regression and Cox proportional hazards regression models is motivated by the prevalent use of these methods in clinical applications.

4.5 Evaluation

We evaluated our research on ECG data from patients in the DISPERSE2 [21] and MERLIN-TIMI36 [73] trials (refer to Appendix A.1 for details). We used data from the DISPERSE2 trial to discover high risk patterns, and tested these patterns on data from the MERLIN-TIMI36 trial. We used the incidence of CVD over a follow-up period of 90 days as the endpoint in both groups.

The DISPERSE2 trial had available data from 765 patients with 14 deaths during follow-up. We excluded data from 4 of these patients who died due to the incidence of myocardial infarction (MI) prior to the mortality event. The MERLIN-TIMI36 trial had available data from 2,302 placebo patients with 57 deaths. Data from 16 of these patients who died was excluded due to the incidence of MI before mortality. For both the DISPERSE2 and MERLIN-TIMI36 patients we used the first 24 hours of ECG recorded after hospitalization for training and testing due to the availability of these signals for all patients, and due to the definition of HRV metrics on 24 hour data (which we subsequently compare our work to). On average, each 24 hour recording in DISPERSE2 contained 103,180 instantaneous heart rate measurements, while in MERLIN-TIMI36 the average length of the heart rate sequences was 102,710.

Since both the DISPERSE2 and MERLIN-TIMI36 trials contained the timing of events and the censoring times for all patients within the 90 day follow-up period, we employed the pattern discovery formulation using the C-index to assess patterns. We further searched for patterns of lengths 6, 8, 10, 12 and 14 (consistent with earlier studies [105]) with $W = 5$ and $d = 2$ under the l_1 norm, and combined the top 5 patterns for each length into a Cox proportional hazards regression model (described in detail in Appendix B.1) developed on the DISPERSE2 data. This model was then applied to data from the MERLIN-TIMI36 trial for testing. All parameters, including W and d were chosen on the training data with testing on the MERLIN-TIMI36 data being carried out blinded to endpoints.

We used Kaplan-Meier survival analysis to compare the mortality rates for patients partitioned into high and low risk groups by our heart rate-based model. This was done by estimating the hazard ratio and 95% confidence interval (CI) for the predictions made by our model for the endpoint of death over a 90 day follow-up. The categorization of patients into high or low risk groups for this analysis was performed by dichotomizing the predictions at the highest quartile consistent with earlier studies to evaluate methods for risk stratification in the setting of ACS [93]. We studied the heart rate-based model on univariate analysis, and also in multivariate models that additionally included the HRV metrics proposed for risk stratification by the Task Force of the European Society of Cardiology and the North American Society of

Pacing and Electrophysiology (Table 2.1). All HRV metrics were dichotomized at the highest quartile consistent with the dichotomization of the predictions of the heart rate-based model.

4.6 Results

4.6.1 Univariate Results

Results of univariate analysis for the predictions made by the model based on heart rate patterns discovered in the DISPERSE2 data , hart rate pattern discovery (HRPD)-Model, and the HRV metrics are presented in Table 4.1. HRPD-Model showed a statistically significant (i.e., $p < 0.05$) association with the endpoint of death in the MERLIN-TIMI36 study. The result in Table 4.1 can be interpreted as roughly a three- to four-fold increased risk of death per unit time in patients found to be at high risk by the pattern discovery-based model. Of the other metrics, HRV-SDANN, HRV-ASDNN and HRV-LF/HF were also significantly associated with death during follow-up. The highest hazard ratio of all these metrics was observed for HRV-LF/HF, followed by HRPD-Model. In general, our results parallel earlier findings in the clinical literature showing that information in HRV can identify patients at an elevated risk of death following ACS.

The Kaplan-Meier mortality curve for HRPD-Model is presented in Figure 4.1. Patients classified as being at high risk by HRPD-Model were at a consistently elevated risk of death during the entire 90 day period following ACS.

4.6.2 Multivariate Results

Table 4.2 presents the correlation between the predictions of HRPD-Model and the HRV metrics. HRPD-Model had low to moderate correlation with the different HRV metrics, suggesting that the results of this approach can be usefully combined with

Table 4.1: Univariate association of predictions with 90 day death in the MERLIN-TIMI36 dataset (HRPD-Model = heart rate pattern discovery model derived from DISPERSE2 data).

Method	Hazard Ratio	P Value	95% CI
HRV-SDNN	1.56	0.275	0.70-3.43
HRV-SDANN	2.33	0.021	1.14-4.78
HRV-ASDNN	2.22	0.034	1.06-4.65
HRV-RMSSD	0.95	0.918	0.36-2.46
HRV-PNN50	0.99	0.978	0.38-2.55
HRV-HRVI	1.82	0.122	0.85-3.91
HRV-LF/HF	3.54	<0.001	1.80-6.97
HRPD-Model	3.46	<0.001	1.76-6.78

Table 4.2: Pearson correlation between HRDP-Model predictions and HRV in the MERLIN-TIMI36 dataset (HRPD-Model = heart rate pattern discovery model derived from DISPERSE2 data).

SDNN	SDANN	ASDNN	RMSSD	PNN50	HRVI	LF/HF
0.14	0.15	0.13	0.00	-0.02	0.13	0.34

Table 4.3: Multivariate association of predictions with 90 day death in the MERLIN-TIMI36 dataset (HRPD-Model = heart rate pattern discovery model derived from DISPERSE2 data).

Method	Hazard Ratio	P Value	95% CI
HRV-SDNN	0.33	0.142	0.08-1.45
HRV-SDANN	2.64	0.078	0.90-7.75
HRV-ASDNN	2.46	0.091	0.87-6.99
HRV-RMSSD	0.71	0.632	0.18-2.86
HRV-PNN50	1.18	0.808	0.31-4.47
HRV-HRVI	1.05	0.936	0.33-3.34
HRV-LF/HF	2.16	0.057	0.98-4.76
HRPD-Model	2.28	0.031	1.08-4.82

Table 4.4: Comparison of pattern discovery runtime (in seconds) for brute force (BF), max-min clustering-based (MM) and our greedy centroid selection/LSH-based (GL) algorithms.

Pattern Length	BF	MM	GL
$L = 6$	83	76	33
$L = 8$	1,460	111	67
$L = 10$	168,796	1,754	1,061
$L = 12$	>3 days	29,058	9,300
$L = 14$	>3 days	78,631	>3 days

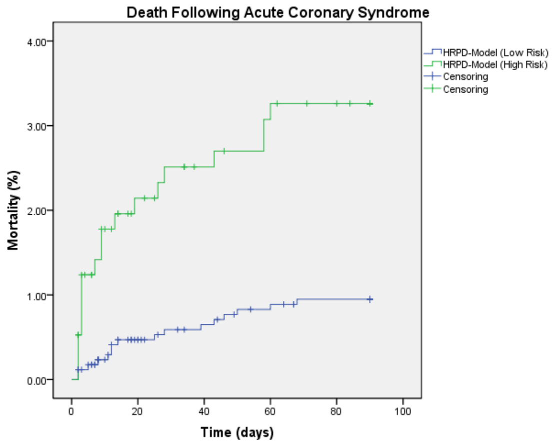


Figure 4.1: Mortality rate in patients categorized as high and low risk by HRPD-Model (HRPD-Model = heart rate pattern discovery model derived from DISPERSE2 data).

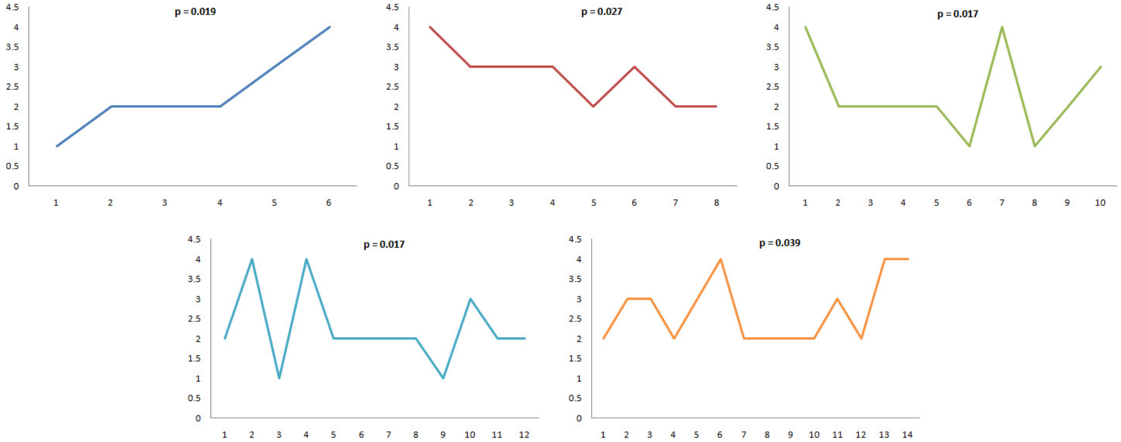


Figure 4.2: Significant heart rate patterns within HRPD-Model (HRPD-Model = heart rate pattern discovery model derived from DISPERSE2 data).

the results of existing metrics.

On multivariate analysis (Table 4.3), HRPD-Model was an independent predictor of death during follow-up, even after adjusting for information in other heart rate-based metrics. None of these other metrics showed an association with the endpoint in the multivariate model at the 5% level (i.e., $p < 0.05$) although the weak association of HRV-SDANN, HRV-ASDNN and HRV-LF/HF at the 10% level suggests that these metrics may have shown a stronger association on a larger dataset.

The complementary nature of the information provided by HRPD-Model to the HRV metrics was further seen on comparison of the C-index of multivariate models with only the HRV metrics included (0.700) to the C-index of multivariate models with both the HRV metrics and HRPD-Model predictions included (0.746). Consistent with the results in Table 4.3, the addition of HRPD-Model to the HRV metrics improved discrimination between high and low risk patients.

4.6.3 Significant Heart Rate Patterns

The heart rate patterns independently associated with death within HRPD-Model (i.e., the patterns with $p < 0.05$ in the Cox proportional hazards regression model used for HRPD-Model) are shown in Figure 4.2. While an interpretation of these patterns is beyond the scope of this chapter, these data suggest that the results in Section 4.6.1 and 4.6.2 are due to information in patterns of varying lengths discovered by our randomized hashing- and greedy centroid selection-based algorithm.

4.6.4 Computational Efficiency

Empirical comparison of our pattern discovery algorithm with both a brute force algorithm (i.e., that takes time quadratic in the number of unique subsequences U_i in the training data) and a max-min clustering algorithm (that uses a greedy search for centroids similar to our approach but does not use LSH) is presented in Table 4.4.

The use of both greedy centroid selection and LSH greatly improved the runtime of the pattern discovery process.

We note that in our work we carefully chose the hash functions and the parameters of LSH such that the probability of finding all approximate patterns is bounded to be higher than 95%. As a result, even though our algorithm sacrificed accuracy to improve computational efficiency relative to (say) a brute force approach, the theoretical guarantees provided by LSH and our results show that the final predictive power of the models remained high.

4.7 Conclusion

In this chapter, we explored the development of novel computational biomarkers to risk stratify patients for death following coronary attacks. Our biomarkers are based on patterns of heart rate changes discovered from large volumes of historical ECG data both from patients who died in the months immediately following ACS and those that remained event free. Discovering such predictive heart rate patterns is made challenging by the computational demands of pattern discovery, variations across patients, and the need to identify activity that may occur in an approximate form due to noise and the stochastic nature of many physiological phenomena.

We address this goal through a randomized hashing- and greedy centroid selection-based algorithm, coupled with the use of SAX to re-express heart rate time series as symbolic sequences. We refine ideas from our previous research to achieve this [102, 105]. The results of our evaluation on over 3,000 patients show that our algorithm can find heart rate patterns that can be combined in risk stratification models to identify patients at an elevated risk of death. Moreover, the information in the patterns discovered by our algorithm can complement information in other heart rate-based metrics, in particular, time-domain and frequency-domain HRV metrics. This has the potential to advance clinical decision-making for a disease that continues to impose an immense burden globally.

We conclude with a discussion of some limitations of our study and opportunities for future improvements. We believe that more research is needed to supplement the statistical significance of the patterns found by our algorithm with an actual physiological interpretation of these patterns. Our algorithm also showed an increasing runtime as the length of patterns increased, despite the use of both randomized hashing and greedy centroid selection. While this is expected due to the increasing number of unique subsequences for large pattern lengths, this limitation prevented us from searching for very long patterns ($L >> 14$). The use of parallelization can address this issue, but more investigation is needed to determine if information in extremely long patterns can aid risk stratification. Moreover, related to pattern length, we also note that our current approach does not exploit the redundancy in work across increasing pattern lengths. Potentially, the information for smaller pattern lengths can be used in a hierarchical manner to focus the pattern discovery search. Finally, we would like to explore a more disciplined way of learning the definition for approximate patterns which would help discover better represented motifs.

CHAPTER V

Improving Models to Stratify Patients

5.1 Introduction

In Chapters III and IV, we focused mainly on improving existing ECG biomarkers or developing new ECG biomarkers. In this chapter, we take this work a step further by studying how to integrate the new biomarkers arising from our work along with existing clinical variables to stratify patients. We note that modeling is often made challenging by the low prevalence of adverse clinical outcomes. Adequately characterizing patient outcomes in the presence of infrequent endpoints requires data from a large number of patients to acquire sufficient positive examples for training. This is expensive, slow, and places unnecessary burden on patients and caregivers.

In this chapter, we address this challenge by jointly leveraging the benefits of both supervised and unsupervised models and propose 1.5 class learning to better stratify patients for adverse outcomes. Specifically, we develop and compare the following frameworks for 1.5 class learning: (1) an algorithm that combines both supervised and unsupervised methods by adding penalties together to form a new joint optimization problem; (2) a transfer learning algorithm that treats supervised and unsupervised methods as tasks that can be transferred; (3) a multi-task algorithm that learns a common hyperplane and task-specific hyperplanes at the same time for both supervised and unsupervised learning tasks; and (4) a 2-class algorithm with the cost parameter chosen to force all negative examples to be included as support vectors.

When evaluated in a real-world setting on a representative population of patients undergoing percutaneous coronary intervention and also on patients undergoing inpatient surgical procedures to predict rare but potentially serious critical care outcomes within 30 days of the procedure, our integrated use of supervised and unsupervised learning significantly improved the discrimination of adverse mortality and morbidity endpoints. This improvement was consistent relative to different conventional algorithms (including cost-sensitive weighting and sampling-based techniques).

The remainder of this chapter is organized as follows. Section 5.2 first describes the challenges of building clinical models in the presence of small datasets and poorly characterized clinical outcomes. Section 5.3 then presents the traditional 2-class and 1-class support vector machine (SVM) formulations. Section 5.4 describes our different approaches to 1.5-class learning using an SVM framework. Section 5.5 details our evaluation methodology and the results of this evaluation are presented in Section 5.6. Finally, we present a geometric interpretation of 1.5-class SVM learning framework in Section 5.7.

5.2 Overview and Previous Work

The challenges of developing models for surgical complications usually stem from existing datasets available for model derivation being small (e.g., thousands to tens of thousands of patients) and suffering from class imbalance. Capturing enough positive examples to robustly train risk stratification algorithms requires collecting data from a large number of patients. This process is slow, expensive, and burdensome to both caregivers and patients. The challenge of collecting this data is further compounded by the multi-factorial nature of many important events, which means that modeling individual pathophysiological processes underlying outcomes requires an even larger number of training examples and increased need for resources.

Traditionally, models to stratify surgical patients have been developed within a *supervised* learning framework. However, supervised learning approaches focus on characterizing the differences between patients who do or do not experience clinical events, and suffer from the lack of sufficient positive (i.e., event) examples for model training when clinical events occur with diminished prevalence. For example, the rate of a wide range of serious complications, ranging from coma to bleeding requiring transfusion was well below 1% in the American College of Surgeons National Surgical Quality Improvement Program (NSQIP) data sampled at over 200 hospital sites [52]. Collecting additional data to address this issue of class imbalance during model training is often infeasible because of delays and expenses to both patients and caregivers. The costs and complexity of collecting extensive data annotated by experts have impeded the spread of well validated and effective healthcare quality interventions [89].

In comparison to these studies, a much smaller body of work has explored the use of unsupervised approaches for clinical risk stratification. In contrast to existing methods, which attempt to develop models for individual diseases using *a priori* knowledge or labeled training data, this work attempts to identify high-risk patients as anomalies in a population (i.e., as patients lying in sparse regions of the feature space). A few notable studies have explored the use of unsupervised anomaly detection in different clinical contexts. Hauskrecht et al [42] described a probabilistic anomaly detection method to detect unusual patient-management patterns and identify decisions that are highly unusual with respect to patients with the same or similar conditions. Tarassenko et al. [107] applied novelty detection to the problem of detecting masses in mammograms. Campbell et al. [15] showed that this approach could identify blood samples in a population corresponding to patients with rare genetic diseases. Roberts et al. [86] demonstrated the successful use of novelty detection for epileptic seizure detection. Laurikkala et al. [58] similarly investigated the value of novelty detection on vestibular data. Most relevant to our current work, Syed et al [4,5] developed and explored the notion of identifying patients at increased risk of different adverse outcomes using a uniform approach where patients were compared to the rest of the population. In this setting, patients were declared to be at high risk if they were highly dissimilar from other individuals in the population.

In this study, we build upon these earlier results using unsupervised risk stratification. A limitation of this earlier work is that a fully unsupervised approach does not exploit any of the information available in labeled examples. While the absence of a large number of positive examples makes it difficult for supervised models to generalize, we nevertheless believe that there may be useful information in the patient labels beyond the support of these patients in the feature space that can be exploited. Most notably, a key limitation of unsupervised risk stratification is that paradoxically it considers both the healthiest and unhealthiest individuals in a population as being at highest risk (since these examples are most likely to manifest as tails of the patient population). By using label information for patients, we believe that it may be possible to encode the directionality of the anomalies in the unsupervised risk stratification process so that the categorization of risk can be focused mainly on anomalies in the unhealthy direction. We exploit this observation and explore the idea of jointly leveraging the benefits of both supervised and unsupervised risk stratification by proposing 1.5-class learning framework which will be described in detail in later sections.

5.3 Background

Before going into the details of 1.5-class learning, we first present background on 2-class and 1-class SVM learning.

5.3.1 2-Class SVM Classification

Binary or 2-class SVM [109] focuses on learning a hyperplane in a high-dimensional feature space that can be used for classification. Given a training set $\{(\mathbf{x_i}, y_i)|\mathbf{x_i} \in \mathbb{R}^m, y_i \in \{+1, -1\}\}_{i=1}^n$ the soft margin SVM formulation aims to solve the following constrained optimization problem:

$$\begin{aligned} \min_{\mathbf{w},\boldsymbol{\xi}} \quad & \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{n}\xi_i \\ \text{s.t.} \quad & y_i(\mathbf{w}^T\phi(\mathbf{x_i}) - b) \geq 1 - \xi_i \quad \forall i = 1, \ldots, n \\ & \xi_i \geq 0 \end{aligned}$$

where ϕ is a kernel function that maps data into some feature space, and the constant C reflects the cost of misclassification and the ξ_i correspond to the slack variables of the soft margin SVM. The dual form of the problem is given by:

$$\begin{aligned} \min_{\boldsymbol{\alpha}} \quad & \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \mathbb{K}(i,j) - \sum_{i=1}^{n}\alpha_i \\ \text{s.t.} \quad & 0 \leq \alpha_i \leq C \quad \forall i = 1, \ldots, n \\ & \sum_{i=1}^{n}\alpha_i y_i = 0 \end{aligned}$$

where $\mathbb{K}(i,j) = \phi(\mathbf{x_i})^T\phi(\mathbf{x_j})$ is the kernel matrix [91]. The final classification rule for predicting the label of a new example $\mathbf{x}$ is then given by $\hat{y} = \text{sgn}(\mathbf{w}^* \cdot \phi(\mathbf{x}) - b)$, and $\mathbf{w}^* = \sum_{i=1}^{n}\alpha_i y_i \phi(\mathbf{x_i})$ can be obtained by solving the dual formulation.

5.3.2 1-Class SVM Classification

The 1-class SVM [92] aims to estimate the support S of a high-dimensional distribution such that the probability that a point drawn from the input space lies outside S is low. Roughly speaking, in contrast to the 2-class SVM algorithm, which separates two classes in the feature space by a hyperplane, the 1-class SVM attempts to separate the entire dataset from the origin. Given training data of the form $\{(\mathbf{x_i})|\mathbf{x_i} \in \mathbb{R}^d\}_{i=1}^n$ (i.e., with the class labels either not available or ignored for training in an unsupervised setting), the 1-class SVM solves the following quadratic problem (which penalizes feature vectors not separated from the origin, while simultaneously trying to maximize the distance of this hyperplane from the origin):

$$\begin{aligned} \min_{\mathbf{w},\boldsymbol{\psi},\rho} \quad & \frac{1}{2}||\mathbf{w}||^2 - \rho + C\sum_{i=1}^{n}\psi_i \\ \text{s.t.} \quad & \mathbf{w}^T\phi(\mathbf{x_i}) - \rho \geq (-\psi_i) \quad \forall i = 1, \ldots, n \\ & \psi_i \geq 0 \end{aligned}$$

where the constant C expresses the tradeoff between incorporating outliers that are not separated from the origin and minimizing the support region. The dual form of the 1-class SVM problem is:

$$\begin{aligned} \min_{\boldsymbol{\alpha}} \quad & \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j\mathbb{K}(i,j) \\ \text{s.t.} \quad & 0 \leq \alpha_i \leq C \quad \forall i = 1, \ldots, n \\ & \sum_{i=1}^{n}\alpha_i = 1 \end{aligned}$$

The classification rule for predicting whether a new example $\mathbf{x}$ lies within the region of high probability is then given by $\hat{y} = \text{sgn}(\mathbf{w}^* \cdot \phi(\mathbf{x}) - \rho)$ with $\hat{y} \leq 0$ denoting the detection of an outlier, and $\mathbf{w}^* = \sum_{i=1}^{n}\alpha_i\phi(\mathbf{x_i})$ can be obtained by solving the dual formulation.

5.3.3 2-Class SVM Classification under Class-Imbalance

Extensive literature has focused on the use of machine learning to diagnose and prognosticate patients under clinical settings where class is highly imbalanced. Much of this research seeks to develop algorithms for risk assessment through supervised learning applied to historical data [54]. While not specific to clinical applications, different solutions have been proposed to address the issue of class imbalance in a general setting [43], with sampling methods [25] and cost-sensitive methods for imbalanced learning [33] being particularly popular.

5.4 Methods

Here, we describe our proposed methods under the 1.5-class learning framework.

5.4.1 1.5-Class SVM Classification

The first approach is described as the basic 1.5-class classification, where an SVM algorithm is designed to improve performance relative to both supervised (i.e., binary or 2-class SVM) and unsupervised (i.e., 1-class SVM) methods for predicting adverse events. This work differs from the typical problem formulation of semi-supervised learning, in that we do not make use of additional unlabeled data to augment supervised learning [78].

Given the training set $\{(\mathbf{x_i}, y_i) | \mathbf{x_i} \in \mathbb{R}^d, y_i \in \{+1, -1\}\}_{i=1}^n$, we integrate penalties for the both formulations, and minimize the objective function below while satisfying the constraints associated with both 2-class and 1-class SVM training.

$$\begin{aligned}
\min_{\mathbf{w},\boldsymbol{\psi},\boldsymbol{\xi},b,\rho} \quad & \frac{1}{2}||\mathbf{w}||^2 - \rho + C_1 \sum_{i=1}^{n} \psi_i + C_2 \sum_{i=1}^{n} \xi_i \\
\text{s.t.} \quad & \mathbf{w}^T \phi(\mathbf{x_i}) - \rho \geq (-\psi_i) \quad \forall i = 1, \ldots, n \\
& y_i(\mathbf{w}^T \phi(\mathbf{x_i}) - b) \geq 1 - \xi_i \quad \forall i = 1, \ldots, n \\
& \psi_i \geq 0 \\
& \xi_i \geq 0
\end{aligned}$$

where C_1 and C_2 denote the costs assigned to penalties for the 1-class and 2-class SVM problems. Since the formulation above represents a summation of quadratic problems the overall training problem remains quadratic. The dual formulation of the 1.5-class SVM problem is then given below:

$$\begin{aligned}
\max_{\boldsymbol{\alpha},\boldsymbol{\gamma}} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} (\alpha_i y_i + \gamma_i)(\alpha_j y_j + \gamma_j) K(\mathbf{x_i}, \mathbf{x_j}) \\
\text{s.t.} \quad & 0 \leq \gamma_i \leq C_1 \quad \forall i = 1, \ldots, n \\
& 0 \leq \alpha_i \leq C_2 \quad \forall i = 1, \ldots, n \\
& \sum_{i=1}^{n} (\alpha_i y_i + \gamma_i) = 1
\end{aligned}$$

The approach above presents different choices for a decision boundary. For example, setting the decision boundary to $\mathbf{w}^T \phi(\mathbf{x}) - \rho$ can be interpreted as a modification of the 1-class SVM result, while a decision boundary corresponding to $\mathbf{w}^T \phi(\mathbf{x}) - b$ can similarly be considered as an extension of the 2-class SVM case. While these choices may be potentially interesting, for the scope of this study, we make use of a common decision boundary that uses the assignment $b = \rho$ for the optimization problem.

5.4.2 1.5 Class SVM Classification with Transfer Learning

The second approach is done by adopting an approach that can be considered as a specialized case of transfer learning. Transfer learning [81] is typically used to address situations where the data for model training and model application are drawn from different distributions. In such cases, transfer learning provides a way to refine a model between training and application. In our work, we adopt a different use of transfer learning: instead of using transfer learning to refine a model across datasets, we use the same underlying principles to refine a model developed using supervised learning for use in unsupervised learning (and vice versa). In this way, our approach uses transfer learning to refine a model across tasks or problem formulations rather than datasets while modeling surgical complications.

Popular methods for transfer learning include instance-based, feature-based, and parameter-based algorithms. In instance-based transfer learning, the goal is to address the issue of training and test distribution differences by reweighting data in the source domain for use in target domain. Feature-based transfer learning finds a good feature representation that is common to both source and target domains. Since both instance-based and feature-based transfer learning are essentially focused on the situation of refining a model between datasets, they are not relevant to our work. Instead, the approach of parameter-based transfer learning, which focuses on transferring parameters between similar but distinct tasks is most relevant to the clinical problem considered here.

We propose a transfer learning extension to 2-class and 1-class SVM classification. Given the training set $\{(\mathbf{x_i}, y_i)|\mathbf{x_i} \in \mathbb{R}^d, y_i \in \{+1, -1\}\}_{i=1}^{n}$, we first utilize 2-class SVM classification for finding a maximum margin boundary $\mathbf{w_2^*}$. Our transfer learning formulation then transfers this 2-class boundary to the 1-class SVM task by solving the following optimization problem:

$$\begin{aligned}\min_{\mathbf{w},\boldsymbol{\psi},\rho} \quad & \frac{1}{2}||\mathbf{w}-\mathbf{w_2^*}||^2-\rho+C\sum_{i=1}^{n}\psi_i \\ \text{s.t.} \quad & \mathbf{w}^T\phi(\mathbf{x_i})-\rho\geq(-\psi_i) \quad \forall i=1,\ldots,n \\ & \psi_i\geq 0\end{aligned}$$

This model regularizes the 1-class SVM solution $\mathbf{w}$ towards the model parameter $\mathbf{w_2^*}$ obtained from the 2-class SVM classification task instead of regularizing $\mathbf{w}$ by itself. In this setting, the regularization term C expresses the tradeoff between slacks and the distance between the transferred model and original model. The model learned will generally be closer to the 2-class SVM task model parameter $\mathbf{w_2^*}$ when C has small values.

Similar to other SVM formulations, solving the dual of this optimization problem is more convenient and provides the advantage of using the kernel trick. In the interest of space, we only present the dual formulation and omit the derivation process (which can be easily done by introducing the Lagrangian). Also, since we already know from section 5.3.1 that $\mathbf{w_2^*}=\sum_{i=1}^{n}\alpha_i^*y_i\phi(\mathbf{x_i})$, thus the dual form of the 2-to-1 SVM transfer problem can be written as:

$$\begin{aligned}\min_{\boldsymbol{\alpha}} \quad & \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j\mathbb{K}(i,j)+\mathbf{w_2^*}\sum_{i=1}^{n}\alpha_i\phi(\mathbf{x_i}) \\ & =\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j\mathbb{K}(i,j)+\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j^*y_j\mathbb{K}(i,j) \\ \text{s.t.} \quad & 0\leq\alpha_i\leq C \quad \forall i=1,\ldots,n \\ & \sum_{i=1}^{n}\alpha_i=1\end{aligned}$$

The transfer learning algorithm can also be applied to first find the one-class SVM boundary $\mathbf{w_1^*}$, and then transfer it using 2-class SVM. The primal formulation is given here:

$$\begin{aligned}
\min_{\mathbf{w},\boldsymbol{\xi}} \quad & \frac{1}{2}||\mathbf{w} - \mathbf{w_1^*}||^2 + C\sum_{i=1}^{n}\xi_i \\
\text{s.t.} \quad & y_i(\mathbf{w}^T\mathbf{x_i} - b) \geq 1 - \xi_i \quad \forall i = 1, \dots, n \\
& \xi_i \geq 0
\end{aligned}$$

We omit the interpretation here since it follows naturally from the previous discussion when transferring from two-class SVM to one-class SVM. Since we already know from section 5.3.2 that $\mathbf{w_1^*} = \sum_{i=1}^{n} \alpha_i^* \phi(\mathbf{x_i})$, the dual form of the 1-to-2 SVM transfer problem is given by:

$$\begin{aligned}
\min_{\boldsymbol{\alpha}} \quad & \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \mathbb{K}(i,j) + \mathbf{w_1^*}\sum_{i=1}^{n}\alpha_i y_i \phi(\mathbf{x_i}) - \sum_{i=1}^{n}\alpha_i \\
= & \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \mathbb{K}(i,j) + \sum_{i=1}^{n}\alpha_i\big(\sum_{j=1}^{n}\alpha_j^* y_i \mathbb{K}(i,j) - 1\big) \\
\text{s.t.} \quad & 0 \leq \alpha_i \leq C \quad \forall i = 1, \dots, n \\
& \sum_{i=1}^{n}\alpha_i y_i = 0
\end{aligned}$$

5.4.3 1.5 Class SVM Classification with Multitask Learning

Multi-task learning [34] is used in scenarios where there are relations between tasks to learn. By exploiting shared structure between these tasks the learning process can be improved. In our work, we build upon this general principle and explore a modification to traditional multi-task learning. Specifically, in contrast to the more usual approach of improving the ability to learn *multiple* models through multi-task learning our focus is on leveraging shared structure between supervised and unsupervised learning to improve the *single* task of stratifying patients.

We propose a multi-task learning extension to 2-class and 1-class SVM classification by simultaneously learning these two related tasks. We start by assuming that the common hyperplane across tasks is $\mathbf{w}_0$ and the task-specific hyperplanes are $\mathbf{v}_1$ and $\mathbf{v}_2$ respectively for the 1-class and 2-class tasks. The regularization parameters $C1$ and $C2$ can be varied depending on the notion of how common the two tasks are. The optimization problem can then be formulated as follows:

$$\begin{aligned}
\min_{\mathbf{w}_0,\mathbf{v}_1,\mathbf{v}_2,\xi,\eta} \quad & \sum_{i=1}^{n} \xi_i + \sum_{i=1}^{n} \eta_i + C_1(\|\mathbf{v}_1\|^2 + \|\mathbf{v}_2\|^2) + C_2\|\mathbf{w}_0\|^2 \\
\text{s.t.} \quad & y_i((\mathbf{w}_0 + \mathbf{v}_2)^T \mathbf{x}_i - b) \geq 1 - \eta_i \quad \forall i = 1, \ldots, n \\
& (\mathbf{w}_0 + \mathbf{v}_1)^T \mathbf{x}_i - \rho \geq -\xi_i \quad \forall i = 1, \ldots, n \\
& \eta_i \geq 0 \quad \forall i = 1, \ldots, n \\
& \xi_i \geq 0 \quad \forall i = 1, \ldots, n
\end{aligned}$$

In the dual, we assume a multiplicative factor of 1/2 for the C_1 and C_2.

$$
\begin{aligned}
\max_{\alpha,\gamma} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2C_2} \sum_{i=1}^{n} \sum_{j=1}^{n} (\alpha_i y_i + \gamma_i) \mathbb{K}(\mathbf{x_i}, \mathbf{x_j}) (\alpha_j y_j + \gamma_j) \\
& - \frac{1}{2C_1} \sum_{i=1}^{n} \sum_{j=1}^{n} \gamma_i \mathbb{K}(\mathbf{x_i}, \mathbf{x_j}) \gamma_j \\
& - \frac{1}{2C_1} \sum_{i=1}^{n} \sum_{j=1}^{n} (\alpha_i y_i) \mathbb{K}(\mathbf{x_i}, \mathbf{x_j}) (\alpha_j y_j) \\
\text{s.t.} \quad & 0 \leq \alpha_i \leq 1 \quad \forall i = 1, \ldots, n \\
& 0 \leq \gamma_i \geq 1 \quad \forall i = 1, \ldots, n \\
& \sum_{i=1}^{n} (\alpha_i y_i) = 0 \\
& \sum_{i=1}^{n} \gamma_i = 0
\end{aligned}
$$

Adding $-\rho$ to the primal objective function has the effect of setting the sum of γ equal to 1 instead of 0 in the dual constraint. Solving the optimization problem above, we can reconstruct our hyperplanes through the following equations with $\mathbf{w}_0$ + $\mathbf{v}_1$ corresponding to the resulting hyperplane found for the 1-class task and $\mathbf{w}_0$ + $\mathbf{v}_2$ corresponding to the resulting hyperplane for the 2-class task.

$$
\begin{aligned}
\text{Hyperplane:} \quad \mathbf{w_0} &= \frac{1}{C_2} \sum_{i=1}^{n} (\alpha_i y_i + \gamma_i) \phi(x_i) \\
\mathbf{v_1} &= \frac{1}{C_1} \sum_{i=1}^{n} (\gamma_i) \phi(x_i) \\
\mathbf{v_2} &= \frac{1}{C_1} \sum_{i=1}^{n} (\alpha_i y_i) \phi(x_i)
\end{aligned}
$$

5.4.4 2-Class SVM Classification with $C \to 0$

Our final approach to 1.5-class learning is based on the traditional 2-class problem. We observe that using extremely low values of the cost parameter C (e.g., 2^{-30}) for the 2-class problem affects the decision hyperplane in ways that parallel 1.5-class learning by forcing all negative examples to behave as support vectors. While we defer a detailed interpretation of this approach to Section 5.7, the basic idea associated with our fourth and final approach for 1.5-class learning is to leverage the traditional 2-class problem with values of C that are close to zero.

5.5 Evaluation

Consistent with our goal of predicting cardiac events, we evaluated our algorithm on cardiac data from the DISPERSE2-TIMI33 trial and the MERLIN-TIMI36 trial (described in A.1). Those data were used to develop and validate models to predict CVD and MI. The derivation cohort comprised the DISPERSE2-TIMI33 data and a random sample of the patients in the MERLIN-TIMI36 data (a quarter of the patients in the MERLIN-TIMI36 data, that is, 1640 patients, chosen at random). Within the derivation cohort, the DISPERSE2-TIMI33 data were used for training models to predict CVD and MI, and the MERLIN-TIMI36 data were used for internal selection of model parameters of the different approaches. The remaining 4920 patients in the MERLIN-TIMI36 cohort were used for validation. Predictive models were trained on DISPERSE2-TIMI33 data to predict CVD and MI within 90 days after nonST-elevation ACS and tested in MERLIN-TIMI36 for the endpoints of CVD and MI within 365 days after non-ST-elevation ACS. The difference in training and testing horizons was based on a shorter available follow-up in the DISPERSE2-TIMI33 data set relative to the MERLIN-TIMI36 data set.

To show general applicability of our algorithm, we also used data from the BMC2 multicenter interventional cardiology registry data (described in Appendix A.2) to develop and validate separate models to predict in-hospital complications of percutaneous coronary intervention (PCI). Models were trained on data from 22,023 patients undergoing percutaneous coronary intervention in 2008, with internal selection of model parameters on 18,993 patients undergoing percutaneous coronary intervention in 2007, and validation on 20,289 patients undergoing percutaneous coronary intervention in 2009. The rationale for our decision to separate data across annual boundaries was that a natural use case of any model developed on a training cohort of patients would likely involve applying that model to future patients. The features used during model training corresponded to a mix of patient characteristics, cardiac status, features related to myocardial infarction, comorbidities, pre-procedure laboratory results, contraindications, pre-procedure therapy, cardiac anatomy and function. The complications of PCI explored in this study included: endoscopic coronary artery bypass graft (ECABG), death (DTH), vascular access (VASC), repeated procedure (RP), stroke (STRK), and gastrointestinal bleed (GI).

All of the algorithms described earlier were implemented using the MOSEK software package for large-scale convex programming (http://www.mosek.com). Each model was trained using a linear kernel with cost parameters chosen by cross-validation from the set $2^{[-10,10]}$. The exception was the 2-Class SVM with values of C approximating zero. For this, we chose a different range of $2^{[-30,-20]}$. In addition to comparing those approaches, cost-sensitive weighting was also considered. 2-class SVM models were trained in this case with the cost parameters chosen to assign a weight to positive examples that was inversely proportional to how infrequently they occurred in the data. For the transfer and multi-task-based SVM approaches it is possible obtain two models for each approach (i.e, transferring from 1-class to 2-class, transferring from 2-class to 1-class, 2-class multi-task, 1-class multi-task). We considered each of these models, as well as the best model chosen for each of the transfer learning and multi-task learning approaches chosen by cross-validation.

For a given SVM model, each patient was assigned a risk score defined as the distance of the patients' predicted score from the decision boundary. We assessed the predictive ability of the SVM models by calculating the AUROC for the test patient scores relative to the different endpoints.

5.6 Results

	BMC2						Cardiac	
	ECABG	DTH	VASC	RP	STRK	GI	CVD	MI
Event Rate	0.22%	0.09%	2.2%	0.49%	0.23%	0.99%	4.4%	7.5%
1-class	0.616	0.655	0.504	0.560	0.566	0.545	0.758	0.628
2-class	0.631	0.621	0.691	0.660	0.639	0.820	0.690	0.570
2-class (W)	0.710	0.870	0.701	0.662	0.684	0.835	0.746	0.595
1.5-class	0.783	0.895	0.696	0.702	0.736	0.846	0.739	0.618
Transfer (1-2)	0.633	0.614	0.691	0.659	0.639	0.820	0.759	0.628
Transfer (2-1)	0.628	0.768	0.595	0.534	0.558	0.764	0.758	0.625
Transfer	0.628	0.768	0.691	0.659	0.639	0.820	0.758	0.625
multi-task-1	0.584	0.658	0.656	0.671	0.509	0.673	0.757	0.629
multi-task-2	0.633	0.614	0.692	0.674	0.640	0.825	0.690	0.570
multi-task	0.584	0.658	0.692	0.674	0.640	0.825	0.757	0.629
2-class (Low-C)	0.798	0.907	0.695	0.700	0.757	0.852	0.748	0.597

Table 5.1: AUROC values comparison of 2-class, 1-class and several 1.5-class SVM approaches for different adverse outcomes within 30 days in patients undergoing inpatient surgical procedures.

Table 5.1 presents the results of our evaluation. For traditional SVM classification, we observed that the use of the 2-class SVM with cost-sensitive weighting improved performance relative to the use of 1-class and 2-class SVM classification. This improvement was consistent for all eight of the study endpoints across datasets.

For the 1.5-class learning approaches with transfer learning, there was no clear trend in terms of transferring from 1-class to 2-class (or vice versa). The choice of which one performed better was not dependent on any obvious statistical property of the data (e.g., prevalence of complications). Therefore, the 1.5-class learning approach with transfer result is an integrated model that is based on using the best of the 1-to-2 and 2-to-1 models (determined during training). The same goes for 1.5-class learning approach using multi-task learning. However, the improvement between those approaches when compared to traditional SVM classification was not significant.

For the other 1.5-class learning approaches, the 1.5-class SVM and the 2-class SVM with values of C chosen to be close to zero both improved discrimination relative to all of traditional SVM algorithms for five out of the six study endpoints (ECABG, DTH, RP, STRK, GI) in the BMC2 cohort. For each of these five endpoints the improvement was significant. In contrast, for the only remaining endpoints (VASC) where the 2-class SVM performed better relative to these approaches the difference in discrimination was marginal. In general, the 1.5-class SVM and the 2-class SVM with values of C close to zero performed analogously although the 2-class SVM with values of C close to zero achieved slightly better results. However, the effect is not as clear as in the Cardiac dataset.

5.7 Interpretation

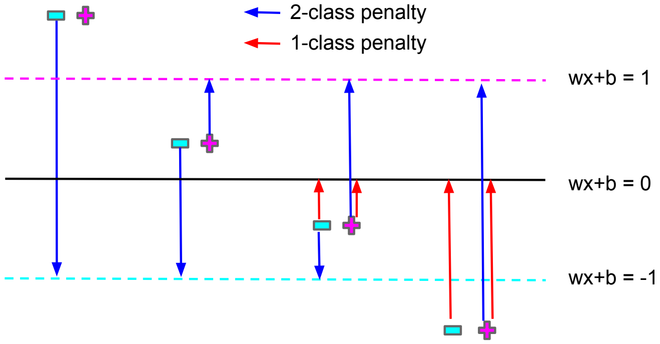


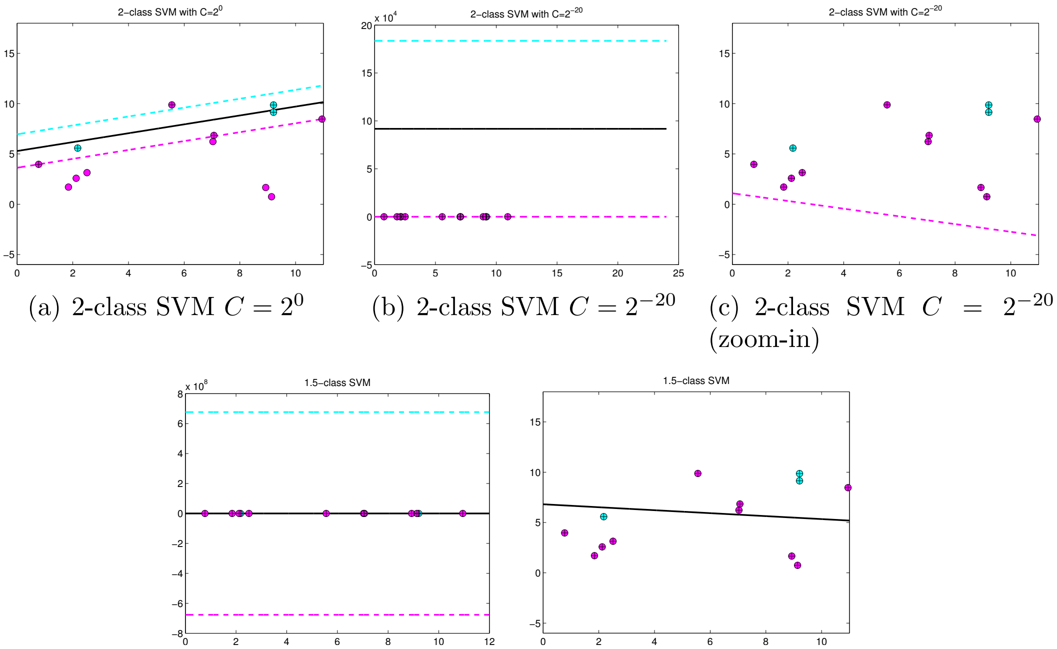
Figure 5.1: Penalization of 1.5-class SVM

We supplement the results presented in Section 5.6 with a brief interpretation. We start by observing that the use of information from both the 2-class and 1-class problems improved clinical assessment for a majority of the endpoints. The improvements in the ability to stratify patients were most prominent for the 1.5-class SVM and the use of 2-class learning with values of C close to zero. In contrast, the transfer learning SVM and multi-task learning SVM did not achieve similar improvements. This results can be attributed in part to the indirect nature of these approaches to achieve a coupling of the 1-class and 2-class problems relative to the other methods. For the transfer learning SVM, the sequential nature of integrating aspects of 1-class and 2-class learning results in isolating the imperfectly characterized training labels from the benefits of support regularization. Similarly, for the multi-task learning SVM, the use of parameter coupling adds a layer of indirection between the 1-class and 2-class SVM formulations. We believe that since the issue of class imbalance results in over-fitting, the maximum pinning down of the decision boundary achieved by the 1.5-class SVM and the 2-class SVM with values of C close to zero is most useful.

The geometric properties for the two best performing methods can be interpreted as follow.

Figure 5.1 shows how the 1.5-class SVM penalizes violations of both 1-class SVM and 2-class SVM constraints. Points on the wrong side of the margins based on labels are penalized as in 2-class SVMs while points below the decision boundary regardless of label are penalized as in 1-class SVMs.

Figures 5.2(a-e) illustrate how the 2-class SVM with extremely low values of the regularization term C effectively becomes a 1.5-class method. We generate synthetic 2 dimensional imbalanced data that is not linearly separable. Panel (a) shows the decision boundary when a moderate value of $C = 2^0$ is used. Note that only the points on or within the margin are the support vectors as is the prototypical view of 2-class SVMs. Panel (b) shows how dramatically the decision boundary changes when $C = 2^{-20}$ is used. By zooming in to the data in Panel (c) we see that all the data points including all the negative data points become support vectors and



(a) 2-class SVM $C = 2^0$

(b) 2-class SVM $C = 2^{-20}$

(c) 2-class SVM $C = 2^{-20}$ (zoom-in)

(d) 1.5-class SVM

(e) 1.5-class SVM (zoom-in)

Figure 5.2: Illustration of comparison between 2-class SVM with low C value versus 1.5-class SVM using synthetic data.

thus help determine the decision boundary. This is a visualization of the crucial insight that makes clear how for extremely small values of C **and** highly imbalanced data, the distribution of the negative data points comes to strongly influence the decision boundary. This is exactly the effect explicitly created through the combined supervised and unsupervised constraints used in the 1.5-class SVM. Indeed, in Panels (d) and (e) we see the result for 1.5-class SVM showing once again that all the data points becomes support vectors.

Figure 5.3 further illustrates the effects of changing the regularization term C. The validation AUROC for the endpoint of ECABG is shown as a function of changing values of C. It is clear that as the value of C is substantially reduced, the AUROC is substantially improved. Consistent with our approach, the value is somewhat flat over a regular range of choices for C and the improvements really manifest at extremely small values of the regularization term, specifically values $C < 2^{-20}$. Note that the AUROC drops to 0.5 when the value of C is reduced further due to numerical issues

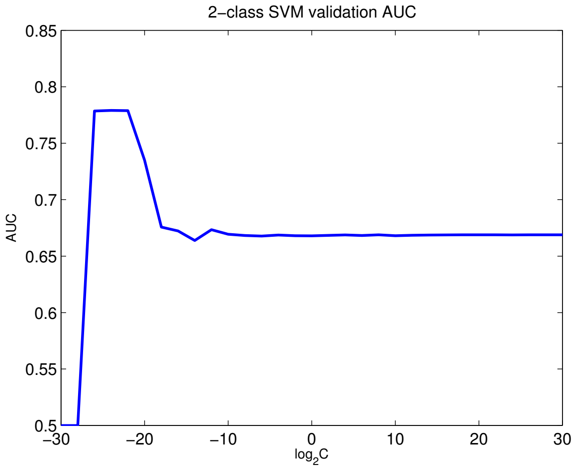


Figure 5.3: 2-class SVM validation AUROC

unrelated to the effect.

5.8 Conclusion

In this study, we considered the problem of class imbalance in clinical datasets. In this setting we explored the idea of 1.5-class models to stratify patients. The focus of our research is to integrate the properties of both 2-class and 1-class models for clinical assessment. At a high level, our approach can be interpreted as either training 2-class models that are regularized for the support of the training data, or training 1-class models where the directionality of anomalies is encoded through available labels. We describe four approaches to implement 1.5-class learning for PCI: (1) a 1.5-class algorithm that combines both supervised and unsupervised methods by adding penalties together to form a new joint optimization problem; (2) a transfer learning algorithm that treats supervised and unsupervised methods as tasks that can be transferred; (3) a multi-task algorithm that learns a common hyperplane and task-specific hyperplanes at the same time for both supervised and unsupervised learning tasks; and (4) a 2-class algorithm with the cost parameter chosen to force all negative examples to be included as support vectors (i.e., C set close to zero).

When evaluated in a representative cohort of patients in the BMC2 registry, the use of 1.5-class learning improved stratification relative to traditional approaches (including the use of cost-sensitive weighting). In particular, the use of the 1.5-class SVM and the 2-class algorithm the C set close to zero both achieved significant improvements relative to traditional SVM algorithms. Geometrically, we observe that both these approaches share similarity. Moreover, unlike the other 1.5-class learning algorithms described in this study they are not affected by potential issues related to the sequential nature of integrating aspects of 1-class and 2-class learning (transfer learning SVM) or the indirect nature of the coupling between problems (multi-task learning SVM).

Future extensions of this work include applying extra regularization parameters in the different 1.5-class algorithms presente (e.g., multi-task learning between the 1-class and 2-class task penalties). We would also like to evaluate our work more rigorously through a variety of kernels. Similarly, another possible direction is to use only the positive examples instead of the whole dataset when dealing with any of the 1-class learning related methods.

We conclude by noting that although we present our research in the setting of complications of PCI, the problem of training clinical models in the presence of generally small datasets and class imbalance is an issue relevant to a broader set of medical applications. In this setting, the improvements obtained through our work may have value in other clinical domains and provide the opportunity to more accurately match patients to treatments and interventions for an extensive set of clinical conditions.

CHAPTER VI

Extension to Atrial Arrhythmias

6.1 Introduction

Our research described in the previous chapters focused on the goal of predicting adverse clinical outcomes in the setting of ACS. As a complement to this work, we consider a different clinical application here, i.e., predicting atrial fibrillation following cardiac surgery. Our aim in this setting is to develop an approach to separate atrial and ventricular parts of the ECG, and to then use information in this source separated ECG to stratify cardiac surgery patients for atrial fibrillation. To achieve this, we propose a novel eigendecomposition algorithm for ECG time-series that leverages information about the underlying cardiac cycle to separate ECG signals into atrial and ventricular components. We then evaluate the clinical utility of MV in atrial components of the ECG to stratify patients for PAF.

In terms of organization, section 6.2 starts by describing the significance of post-operative atrial fibrillation. Then, section 6.3 describes the shortcomings of previous efforts to stratify patients for this endpoint. This is followed by details of our problem formulation for predicting PAF. We then describe how this approach can be implemented. We evaluate our work through a set of experiments on data collected from University of Michigan Cardiovascular Center and finally present and discuss the risk prediction results in section 6.5.

6.2 Significance

PAF occurs in 10% to 65% of the patients undergoing cardiac surgery [114, 80, 66, 7]. In PAF the upper chambers (atria) of the heart tend to fibrillate or contract fast and irregularly, preventing successful emptying of blood into the lower (ventricular) chambers; consequently blood may pool in the heart and clot causing strokes and other morbidities. This causes blood to pool in the heart and clot, producing strokes and other morbidities that increase risk of postoperative mortality [69]. The highest incidence of PAF is typically seen on the second and third postoperative day, with fewer patients developing the condition either in the early postoperative period, or four or more days after surgery [69, 47]. PAF is associated with increased postoperative mortality and morbidity [69]. It also imposes a significant burden on the healthcare system by resulting in longer and more expensive hospital stays [66].

Prophylactic use of beta-adrenergic blockers and amiodarone postoperatively has been shown to reduce the incidence of PAF substantially [72, 39, 40]. However, while considering these treatments and other options such as rhythm control and anti-coagulation, the benefits of therapy must be balanced against adverse effects. For example, blinded therapy in the Prophylactic Amiodarone for the Prevention of Arrhythmias that Begin Early After Revascularization, Valve Replacement, or Repair (PAPABEAR) trial was more likely to be withdrawn in patients treated with amiodarone, largely because of a 3-fold increase in bradycardia requiring pacing and QTc interval prolongation greater than 650 milliseconds [84].

Identifying parameters that are associated with an increased risk for PAF, and provide information complementary to existing tools, may help promote a better and more complete understanding of the pathophysiology of the disorder. This can potentially allow for other therapies to be considered, to treat patient populations more comprehensively. Moreover, identifying patients at an elevated risk of PAF can allow for finer prophylactic administration of pharmacological therapy. This involves both preventing possible side effects in patients who are at low risk and may otherwise receive drugs unnecessarily, and also more aggressive care for high risk patients than they presently receive.

6.3 Overview and Previous Work

A number of different clinical metrics have been proposed to predict PAF. Older age has been shown to be consistently associated with a higher incidence of PAF [114, 70, 46], most likely due to increased atrial fibrosis and dilation in older patients. Large observational studies have also found an association between other clinical characteristics and PAF, although the results of these studies have often been conflicting. Hypertension has been found to predict atrial fibrillation after cardiac surgery [36], possibly due to fibrosis and dispersion of atrial refractoriness [2, 6]. Men also appear to be more likely than women to develop PAF after coronary artery bypass graft (CABG) [114, 2, 6]. It is believed that this effect may be due to differences in ion-channel expression and hormonal effects on autonomic tone. Previous atrial fibrillation and previous congestive heart failure have also shown an association with PAF [70]. In addition, procedural information such as aortic cross-clamp time and location of venuous cannulation have been found in some, but not all, studies to have predictive value for PAF [66]. Postoperative factors such as respiratory compromise and prolonged ventilation have also been suggested [6].

There is also an extensive literature on ECG-based metrics to risk stratify patients for PAF. Most of this work has focused on detecting an abnormal prolongation of P-wave duration on the surface ECG, as a way to identify intra-atrial conduction defects [20]. Various time-domain features (e.g., P-wave duration [20, 9, 5, 30, 24, 95, 53], isoelectric interval duration [20], signal averaged P-wave duration [98, 35, 113, 8, 22], P-wave dispersion [30], P-wave variance [5, 110], P terminal force and spatial velocity [71], and PR interval length [82]) have been explored. The P-waves and PR intervals in surface ECG have also been studied using frequency-domain methods to compute spectral characteristics [96, 97, 44]. In addition to analyses based on the P-wave, HRV has been analyzed to study a potential role of dysfunction of the autonomic heart rate control in inducing PAF [31, 45]. It is believed that a change in sympathetic and vagal tone may precede atrial fibrillation. However, it has been demonstrated that the prediction is feasible just prior to the onset of PAF, which reduces the window of efficient prophylactic therapy significantly [94].

Despite much promise, most previously proposed methods suffer from issues related to inadequate precision and recall for clinical use. The small patient populations (64 to 240 patients) also make it difficult to generalize the findings from these studies to larger population. Other factors that limited the clinical application of ECG-based metrics can be attributed to the lack of fully automated algorithms for some metrics, and the lack of standard definitions (e.g., for P-wave duration). Lack of data regarding specific treatment implications also limits the routinely incorporation into clinical practice.

Our work focuses on developing and assessing novel ECG metrics that can be clinically deployed in a fully-automated setting to identify patients at risk of PAF. To achieve this, our research builds off recent advances in predicting ventricular arrhythmias, which is MV described in section 2.4. In this work, we adopt a similar approach but note that since the ECG is heavily dominated by ventricular activity, it may be more appropriate to decouple ECG signals into atrial and ventricular components that can be separately assessed for morphological variations. Consistent with the hypothesis proposed earlier, we believe that variability in the atrial ECG may reflect specific instability associated with PAF rather than other kinds of arrhythmias.

We can now narrow down the focus of our work to studying MV in atrial activation as a means of stratifying patients for PAF. Since observing atrial activity over the entire cardiac cycle is made difficult by the presence of ventricular activity, our approach requires first extracting the atrial components of the ECG waveform from the surface ECG. Traditional filtering-based methods are insufficient for this task since the ventricular activity is both higher amplitude, and occupies the same frequency ranges as atrial activity. Instead, we plan to formulate the separation of atrial and ventricular activity as a blind source separation problem, where the aim is to extract the atrial component of ECG waveform, that is to say, to separate out atrial activity from ventricular activity and noise.

6.4 Decomposing ECG into Atrial and Ventricular Components

Based on the cardiac cycle, the P-wave of the ECG consists exclusively of atrial activity. Similarly, the T-wave has almost exclusively ventricular activity. In contrast, the QRS-complex has both atrial and ventricular activity (see Figures 2.2(a) and (b) for detail); we will use these facts crucially in developing our approach for extracting atrial activity.

6.4.1 Problem Formulation

The surface ECG measures electrical activity at different parts of the body and follows a linear instantaneous model [77], i.e., each recording of an ECG lead is a weighted linear combination of the atrial and ventricular components. Thus, the source separation problem that we are trying to solve can be viewed as an instance of the cocktail party problem.

More formally, the source signals at time t are represented by a random vector $\mathbf{s}(t) = [s_1(t), s_2(t), \cdots, s_n(t)]^T \in \mathbb{R}^{n\times 1}$. The observed signals at time t are represented by a random vector $\mathbf{x}(t) = [x_1(t), x_2(t), \cdots, x_m(t)]^T \in \mathbb{R}^{m\times 1}$. Each source signal $s_i(t)$ is a linear combination of the observed signals $\mathbf{x}(t)$ at each time point according to the linear instantaneous model. Our goal is to estimate $\mathbf{s}(t)$ from $\mathbf{x}(t)$ where $\mathbf{s}(t) = \mathbf{W} * \mathbf{x}(t)$ by estimating the unmixing matrix $\mathbf{W} \in \mathbb{R}^{n \times m}$. Note that $\mathbf{W}$ is a constant over all t. In our work, we learn a unmixing matrix that when applied to the recorded ECG signal, $\mathbf{x}(t)$, recovers both atrial and ventricular sources, $s_a(t)$ and $s_v(t)$ respectively.

6.4.2 Existing Method

Most of the existing work on atrial component extraction has focused on surface ECG extracted during atrial fibrillation episodes. The literature suggests that *during atrial fibrillation episodes*, atrial activity consists of small and continuous wavelets (a sawtooth form [23]) with a cycle around 160ms, and therefore has been modeled as a random variable with a distribution described by its histogram, a subgaussian signal. Therefore, based on this model, the atrial signal is said to have negative kurtosis (being subgaussian), while ventricular signal has positive kurtosis (as it is assumed to be supergaussian). When such assumptions hold, i.e., during atrial fibrillation episodes, independent component analysis (ICA), which is capable of extracting independent non-Gaussian sources, has been shown to successfully extract atrial activity.

Our goal, however, is prediction in advance of atrial fibrillation; therefore the assumptions upon which the ICA method is proposed are no longer valid. Nevertheless, for completeness and comparison below we explore the use of ICA for atrial component extraction on ECG that may not contain atrial fibrillation episodes. Specifically, we applied RobustICA, a variant of ICA based on using the kurtosis as contrast function. The component with most positive kurtosis is considered to be the ventricular component, while the one with most negative kurtosis is considered to be the atrial component.

6.4.3 Proposed Method

6.4.3.1 Silence-energy-minimization (SEM)

Our work differs from the standard cocktail party problem in that we have additional *a priori* knowledge of the time frames where *only one* of the speakers is speaking. Specifically, based on cardiac physiology we know that for each source $s_i(t)$ there exists time frames $t \in [a_i, b_i]$ when only that source is active. As described in the previous section and also Figure 2.2, we can see that this is the case for each heart beat: the P-wave is associated exclusively with atrial depolarizaion while the T-wave relates only to ventricular repolarization. Thus there are periods within the ECG when only atrial (P-wave) or ventricular (T-wave) activity is present.

We exploit information about the exclusively atrial or ventricular activity as follows. Let the source signal $\mathbf{s}(t) = [s_a(t), s_v(t)]^T \in \mathbb{R}^{2\times 1}$, where $s_a(t)$ denotes the A-beat (atrial) and $s_v(t)$ denotes V-beat (ventricular) component at a given time point t. Let $S = [\mathbf{s}(t_1), \mathbf{s}(t_2), \cdots, \mathbf{s}(t_T)] \in \mathbb{R}^{2\times T}$ be the whole signal across time over a certain window of multiple heartbeats. Let $X = [\mathbf{x}(t_1), \mathbf{x}(t_2), \cdots, \mathbf{x}(t_T)] \in \mathbb{R}^{m\times T}$ be the entire observed signal across time over the same window of heartbeats, where m represents the number of ECG leads recorded. Note that, in this following derivation, we focus on a single heartbeat for clarity, however in our results we will be looking over windows of multiple beats.

Assuming we can segment out the P- and T-waves in the observed signal, we know that only the ventricular source is active for the duration of P-wave segment $t \in [t_{PS}, t_{PE}]$, and only the atrial source is active for the duration of T-wave segment $t \in [t_{TS}, t_{TE}]$. Therefore, we represent the P-wave (atrial) and T-wave (ventricular) parts of the observed signal as $\mathbf{a}(t)$ and $\mathbf{v}(t)$ respectively, both $\in \mathbb{R}^{m\times 1}$:

$$\mathbf{a}(t) = \begin{cases} \mathbf{x}(t) & \forall t \in [t_{PS}, t_{PE}] \\ \vec{0} & \text{otherwise} \end{cases}$$

$$\mathbf{v}(t) = \begin{cases} \mathbf{x}(t) & \forall t \in [t_{TS}, t_{TE}] \\ \vec{0} & \text{otherwise} \end{cases}$$

Given this, we want to find linear combination vectors $\mathbf{w_a}, \mathbf{w_v} \in \mathbb{R}^{m\times 1}$ where for:

- $t \in [t_{PS}, t_{PE}]$, $s_a(t) = \mathbf{w_a}^T * \mathbf{x}(t) = \mathbf{w_a}^T * \mathbf{a}(t)$
- $t \in [t_{TS}, t_{TE}]$, $s_v(t) = \mathbf{w_v}^T * \mathbf{x}(t) = \mathbf{w_v}^T * \mathbf{v}(t)$

under the following optimization function, where $X_A = [\mathbf{a}(t_1), \mathbf{a}(t_2), \cdots, \mathbf{a}(t_T)]$ and $X_V = [\mathbf{v}(t_1), \mathbf{v}(t_2), \cdots, \mathbf{v}(t_T)]$ both $\in \mathbb{R}^{m\times T}$:

$$\max_{\mathbf{w_a}} ||\mathbf{w_a}^T X_A||^2 - C||\mathbf{w_a}^T X_V||^2 \quad \text{s.t.} \quad ||\mathbf{w_a}||^2 = 1 \tag{6.1}$$

$$\max_{\mathbf{w_v}} ||\mathbf{w_v}^T X_V||^2 - C||\mathbf{w_v}^T X_A||^2 \quad \text{s.t.} \quad ||\mathbf{w_v}||^2 = 1. \tag{6.2}$$

where:

- (6.1) seeks a set of weights $\mathbf{w_a}$ that when applied to a heartbeat's ECG recovers maximal energy in the P-wave region while minimizing the energy in the T-wave region.
- (6.2) seeks a set of weights $\mathbf{w_v}$ that when applied to a heartbeat's ECG recovers maximal energy in the T-wave region while minimizing the energy in the P-wave region.

Given that the P- and T-waves do not overlap in time, (6.1) and (6.2) can be decoupled and dealt with separately. Here, we only derive the solution for (6.1) as the solution for (6.2) is similar. Adding the Lagrangian term, the problem becomes:

$$\max_{\mathbf{w_a}} ||\mathbf{w_a}^T X_A||^2 - C||\mathbf{w_a}^T X_V||^2 - \lambda(||\mathbf{w_a}||^2 - 1)$$

Taking the derivative of the above with respect to $\mathbf{w_a}$ and setting it to zero.

$$X_A^T X_A \mathbf{w_a} - C \cdot {X_V}^T X_V \mathbf{w_a} - \lambda \mathbf{w_a} = 0$$
$$(X_A^T X_A - C \cdot {X_V}^T X_V) \cdot \mathbf{w_a} = \lambda \mathbf{w_a}$$

Therefore, the resultant optimization problem is an eigen-problem, where the solution for $\mathbf{w_a}$ is the eigenvector corresponding to the largest eigenvalue of the following matrix:

$$X_A^T X_A - C \cdot {X_V}^T X_V,$$

where C is a regularization term that controls the degree to which the unwanted ventricular part is attenuated. Similarly, the solution for $\mathbf{w_v}$ is the eigenvector corresponding to the largest eigenvalue of:

$$X_V^T X_V - C \cdot {X_A}^T X_A.$$

Once we've learnt w_a and w_v, we apply it to $\mathbf{x}(t)$ for all time t. This will result in obtaining our estimate of the atrial/ventricular component over the full range of the ECG recording.

6.4.3.2 Extracting P and T-waves

At the core of our approach is identifying the location of the P and T-waves in the ECG beats. Many algorithms have been proposed in the literature to segment cardiac

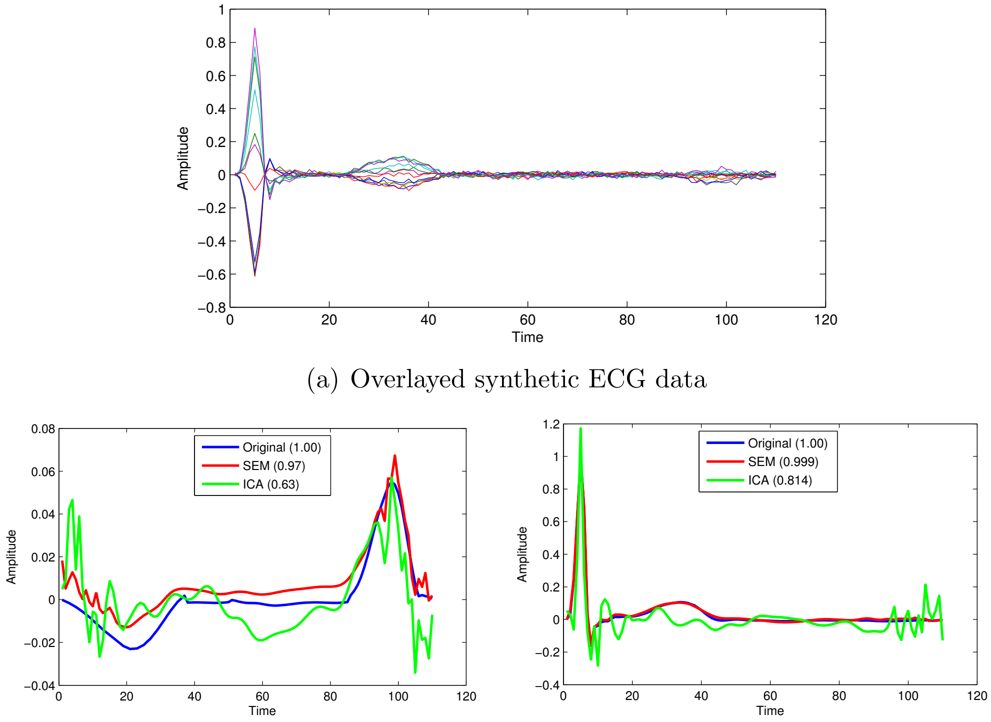
ECG beats into their corresponding P/Q/R/S/T-waves. However, these algorithms are generally unreliable at extracting the P/T-waves in real-world signals due to the relatively small magnitude of the P-wave and subtle changes marking the end of the T-wave. In addition, the real-world data employed in this chapter are especially noisy, due to collection in an operating room (OR) setting, rendering these segmentation algorithms unsuitable for our purposes. As a result of this, we devised a heuristic based on physiology to establish the location of the P and T-waves. Specifically, we attempted to relate the occurrence of these waves to the R-peak, which is the most prominent part of the beat (and therefore the easiest to detect). Our proposed heuristic is as follows: we make the general assumption that there is no ventricular activity (hence atrial part) during 60 to 180 ms before an R-peak, while there is no atrial activity during 80 to 480 ms after an R-peak (hence ventricular part).

Finally, we note that Weisman et al. proposed a method similar to ours to extract atrial electrical activity [111]. However, their method makes an unrealistic assumption that the whole ECG signal can be cleanly segmented into segments of only pure atrial activity only and pure non-atrial activity. Moreover, their method tries to maximize the ratio of energy between atrial part vs non-atrial part, which requires an iterative algorithm when trying to solve the optimization function. Our approach, in contrast, has a closed form solution that is more applicable to real-time systems.

6.5 Experiments and Results

We evaluated our proposed methodology for PAF prediction on two sets of data: synthetic and real-world. We first studied the ability of the two atrial extraction approaches proposed above (i.e., ICA and SEM) to reliably recover atrial and ventricular components on synthetically created ECG data. These experiments were then supplemented by an investigation in a representative real-world clinical cohort of the utility of atrial and ventricular separation in predicting PAF. Details of experiments and results are presented below.

6.5.1 Synthetic Data



(a) Overlayed synthetic ECG data

(b) Original vs. recovered atrial

(c) Original vs. recovered ventricular

Figure 6.1: Comparison of atrial and ventricular components extracted using ICA and SEM on synthetic data ($C = 10$ for SEM).

We created synthetic ECG beats by combining textbook templates of atrial activity (defined as the P-wave and T_A-wave) with ventricular activity (defined as the remaining waves). Specifically, we simulated multi-channel ECG data by using the linear instantaneous model proposed in [77] to combine the atrial and ventricular components with randomly selected weights and additive white Gaussian noise. ICA and SEM were applied to this generated multi-channel ECG to obtain candidate atrial and ventricular components. These components were compared with the ground truth atrial and ventricular activity to assess the ability of ICA and SEM to reliably recover the original signals (using correlation as a performance criteria).

Figure 6.1 presents the results of this experiment. The synthetic multi-channel ECG data created using the approach above is illustrated in Figure 6.1(a). When separated into atrial and ventricular components (Figures 6.1(b) and (c)), the use of SEM for separation provided consistent improvements in the recovery of both atrial and ventricular activity relative to the use of ICA. In particular, the use of prior knowledge about the relative absence of atrial and ventricular activity in SEM yielded a correlation coefficient of greater than 0.97. This was in contrast to the use of ICA, which failed to achieve any reasonable recovery of the atrial component and achieved marginal success dealing with ventricular activity (correlation coefficient of 0.81). Visually, the use of ICA also led to substantially more ripple in the extracted components than the use of SEM.

While we did not rigorously compare the ability of ICA and SEM to separate ECG into atrial and ventricular components on real patient data (owing largely to the absence of known ground truth in real data versus synthetic data), we note than in many cases the use of SEM provided qualitatively better results. For example, as shown in Figure 6.2, the use of SEM on 4-lead ECG data (Figure 6.2(a)) resulted in atrial components with substantially increased energy in the P-wave and PR-interval as opposed to ventricular components with substantially increased energy in the ST-segment and T-wave. This was in contrast to ICA, where the absence of prior knowledge informing such a separation led to a comparatively poorer separation of the signal (Figures 6.2(b) and (c)).

6.5.2 Real-World Data

We supplemented our analysis investigating the abilities of ICA and SEM to separate ECG into atrial and ventricular components with an evaluation of the clinical utility of such a separation on a real-world representative cohort of patients undergoing cardiac surgery. Data from 385 patients undergoing CABG, aortic, or other valvular surgeries

Feature Set	AUROC
Non-ECG	0.66
Non-ECG ECG (No Separation)	0.69
Non-ECG ECG (No Separation) ECG (ICA)	0.70
Non-ECG ECG (No Separation) ECG (SEM)	0.70
Non-ECG ECG (No Separation) ECG (ICA) ECG (SEM)	0.70

Table 6.1: AUROC values for logistic regression models trained using stepwise backward elimination applied to different groups of features.

Feature Set	IDI (p-value)	NRI (p-value)
Non-ECG	0.048 (<0.001)	53.4% (<0.001)
Non-ECG ECG (No Separation)	0.017 (0.026)	25.6% (0.017)
Non-ECG ECG (No Separation) ECG (ICA)	0.004 (0.275)	16.1% (0.091)
Non-ECG ECG (No Separation) ECG (SEM)	Referent	Referent

Table 6.2: IDI and NRI values for logistic regression models trained using stepwise backward elimination applied to different groups of features.

at University of Michigan Hospital (details in Appendix A.3). 90 atrial fibrillation events were annotated. The size of this cohort was considerably larger than previous studies investigating the use of ECG-based metrics to predict PAF (previously a maximum of 240 patients) largely because a focus on exploring a fully-automated approach to predict PAF allowed us to evaluate our approach more rigorously in a larger cohort.

The goal of our investigation was to study the ability of markers based on MV deriving from atrial and ventricular components of the ECG waveform in the OR to predict PAF. We note that since recordings collected once surgery has started are typically too noisy for meaningful analysis, only the first 30 minutes of data in the OR preceding the operation were used. Moreover, since a key question while determining clinical utility is the extent to which any novel markers add information beyond existing variables, we also compared the use of MV measured from atrial and ventricular components of the ECG to baseline clinical features available in the patient EHR (demographics, history and physical exam findings, laboratory reports, and type of surgery) and also based on the unseparated ECG signal (MV measured on each of Leads 1-4).

Specifically, we studied how the discrimination (as assessed by AUROC and IDI, see Appendix B.2) and reclassification (as assessed by the net reclassification improvement, NRI, see Appendix B.3) of logistic regression models trained with different combinations of the features varied. These included logistic regression models trained using stepwise backward elimination applied to: (1) non-ECG features; (2) non-ECG features and features based on the complete ECG signal; (3) non-ECG features and features based on both the complete ECG signal and components derived using SEM; (4) non-ECG features and features based on both the complete ECG signal and components derived using ICA; and (5) non-ECG features and features based on both the complete ECG signal and components derived using both ICA and SEM. In all of these experiments, the stepwise backward elimination process removed one feature during each iteration based on cross-validated results for each step.

Tables 6.1 and 6.2 present the results of this experiment. The inclusion of MV measured from ECG without separation substantially improved performance relative to the use of baseline clinical features by themselves. This performance was further improved with the addition of MV based on atrial and ventricular components derived through both ICA and SEM. The improvement was marginally larger when SEM was used for separation than when ICA was employed. Specifically, we note that when MV markers based on both ICA- and SEM-separated ECG components were included together, the backward stepwise elimination process retained MV based on atrial activity derived using SEM in preference to MV based on all ICA derived components. The IDI and NRI metrics using SEM were also both positive relative to the use of ICA (with NRI showing a 16.1% improvement in reclassification that was significant at the 10% level).

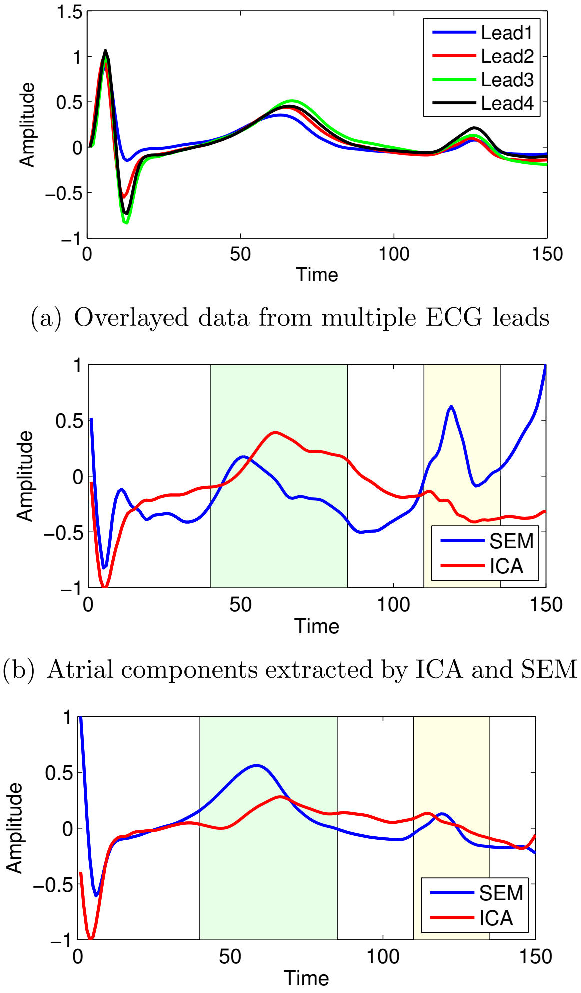
6.6 Conclusion

In this chapter, we focused on the question of developing novel markers that can be used to stratify patients undergoing cardiac surgery for PAF. Given the substantial burden that PAF imposes post-operatively, the ability to identify patients most likely to experience PAF can substantially improve mortality and morbidity (and also reduce healthcare costs) by creating the opportunity to deliver prophylaxis in a timely and personalized manner. The challenge to realizing this, however, is that there are currently no established metrics for PAF risk stratification. To address this need, we explored the development of ECG-based markers in our work that can be deployed in an inexpensive, non-invasive, and fully-automated manner to evaluate patients undergoing cardiac surgery. We focused, in particular, on extending advances in stratifying patients for ventricular arrhythmias (i.e., by quantifying excessive variability in the

ECG waveform) to similarly evaluating the health of the electrical activity of the atria. Central to this is the ability to distinguish lower amplitude atrial activity from higher amplitude ventricular activity. To decompose the ECG into separate components corresponding to both atrial and ventricular activity, we proposed a novel eigen-decomposition approach based on silence energy minimization, which partitions ECG time-series into atrial and ventricular components by exploiting knowledge of the underlying cardiac cycle.

As possible extensions of this work, we believe that the current method can be modified so that more timepoints from the signal (other than just the P- and T-wave) can be utilized. Also, we would like to explore the opportunity of combining the merits of ICA and our SEM method, for example, integrating the *a priori* knowledge that SEM utilizes into ICA through Bayesian learning.

We evaluated our work on both synthetic and real-world data. Our results on synthetic data showed that the use of additional knowledge based on physiology to distinguish between atrial and ventricular activity during the ECG decoupling process substantially improved performance relative to physiology-agnostic approaches such as ICA. When further evaluated on data from a large (for this clinical application) and well-characterized cohort of patients undergoing cardiac surgery, we further observed that the use of physiology to guide ECG separation into atrial and ventricular components achieved better results than the use of a purely statistical approach such as ICA. Moreover, the development of markers based on an analysis of atrial ECG significantly improved models based on baseline clinical features and an assessment of variability within the entire (unseparated) ECG. In particular, our results show that relative to the combination of baseline clinical features and ECG features without separation, our proposed approach can improve classification by over 25% with statistically significant improvements in discrimination. These results have the potential to improve the management of tens of thousands of patients each year.



(a) Overlayed data from multiple ECG leads

(b) Atrial components extracted by ICA and SEM

(c) Ventricular components extracted by ICA and SEM

Figure 6.2: Comparison of atrial and ventricular components extracted using ICA and SEM on actual ECG data ($C = 10$ for SEM). Shaded bands correspond to portions of the cardiac cycle corresponding to ventricular (green) and atrial (yellow) activity.

CHAPTER VII

Conclusion

In this thesis, we proposed several novel computational methods for analyzing large amounts of physiological data. From a technical perspective, our contributions included the development of techniques for computational discovery of new biomarkers from long-term physiological recordings, and for the improved training of classification models for clinical stratification. From a medical perspective, our research spanned both improving patient care following ACS, and being able to deliver prophylaxis in patients undergoing cardiothoracic surgery for atrial arrhythmias.

To reduce the computational complexity of MV for improving existing ECG biomarkers, we presented the concept of adaptive downsampling to reduce the amount of data while retaining the information in rapidly changing parts of the ECG. We described a trace segmentation-based approach to adaptively downsample signals, along with a modified DTW dynamic programming formulation that could leverage these adaptively downsampled inputs. We found an almost 4-fold reduction in runtime relative to DTW, without a significant change in the clinical utility of the MV marker.

We also proposed novel risk markers, advancing existing work on cardiovascular risk stratification through a novel approach that tracks information in short heart rate patterns. We formulated the problem of discovering approximate sequential patterns for risk stratification from large volumes of historical heart rate data. We explored a symbolic transformation of heart rate time series to handle inter-patient variability, and proposed the use of LSH to solve this problem in a computationally efficient manner. A rigorous evaluation of our research on a real-world dataset with long-term ECG signals and detailed follow-ups from over 3,000 patients found that heart rate motifs identify patients at a 2-fold increased risk of death even after adjusting for other clinical metrics.

To improve the learning of risk stratification models, we further proposed a new paradigm called 1.5-class learning that takes advantage of the strengths of both 1-class and 2-class classification models. We explored four separate implementations of the 1.5-class learning idea, and evaluated the relative merits of these different 1.5-class methods on a common dataset when compared to traditional 1-class and 2-class approaches. 1.5-class learning demonstrated genuine clinical utility on several different datasets and across multiple endpoints. We also developed a geometric interpretation for the improvements achieved through 1.5-class learning to help our understanding of the advantages of the approach. In the future, we would like to explore new formulations for 1.5-class learning, as well as to apply it with different non-linear kernels.

To extend the MV risk factor to prediction of atrial arrhythmias, we proposed risk stratification of patients for PAF using information specific to the atrial component of the ECG, and described a new eigendecomposition approach to decomposing ECG signals into atrial and ventricular components. Using this, we measured atrial instability by studying variations in atrial ECG morphology as a means of determining risk for PAF. We rigorously evaluated the hypothesis that separating out the atrial component of the ECG signal can lead to better PAF prediction in a real-world cohort of patients undergoing cardiac surgery. We also compared the relative merits of our approach to atrial component extraction with existing independent component analysis. This work represents the first attempt to develop an integrated electrophysiological assessment of both external (autonomic) and internal (myocardial) factors related to PAF. The data collected during this project will also be used toward creating a public resource that can assist the broader research community in achieving further progress.

Collectively, our efforts showcase the ability of computation to address varied aspects of clinical problems and allow for end-to-end improvements in patient care. As future research within this scope, we hope to extend our work to signals other than ECG (e.g., pulmonary artery pressure or respiration signals). Since the variation of blood pressure and respiratory signals are related to the functioning of the heart, we believe our work can also improve risk stratification when applied to these signals. Another potential extension of our work is to explore ways to leverage multi-channel data. This can provide an opportunity to measure other physiological phenomena (i.e., beyond the kinds of things already included in this thesis) and help us achieve a deeper understanding of the clinical conditions. Finally, while we validated the utility of our approaches on large and representative patient cohorts, closer collaboration with hospitals and doctors to further develop and evaluate these approaches can improve their clinical applicability. By doing this, we believe our methods will have a valuable role in the area of data-driven medicine and be enhanced in their ability to achieve positive societal impact.

APPENDICES

APPENDIX A

Data

A.1 DISPERSE2-TIMI33 and MERLIN-TIMI36

The DISPERSE2-TIMI33 trial [21] enrolled 909 patients if they experienced ischemic symptoms at rest for a duration exceeding 10 minutes with either biochemical marker evidence of MI (defined as Troponin-T, -I, or creatinine kinase-MB elevation greater than the local MI decision limit) or ECG evidence of ischemia (defined as the presence of new or presumably new ST-segment depression ≥0.05 mV, transient ST-segment elevation ≥0.1 mV, or T wave inversion ≥0.1 mV in 2 or more contiguous leads). As part of this study, continuous ECG data were recorded for a median duration of 4 days. Three-lead LifeCard CF Holter monitors were placed within 48 hours of the initial event, and the data were sampled at 128 Hz. Patients were followed up for a period of 90 days for CVD.

The MERLIN-TIMI36 trial [73] studied 6560 patients hospitalized with non-ST-elevation ACS. Patients with moderate- to high-risk clinical features were enrolled within 48 hours of their last ischemic symptoms and treated in a blinded manner with intravenous followed by oral ranolazine or matching placebo. Patients in the

MERLIN-TIMI36 trial received standard medical and interventional therapy according to local practice guidelines and were followed for a median duration of 348 days for CVD and MI. Continuous three-lead 128-Hz ECG recording was initiated at randomization (within 48 hours of the last ischemic discomfort) and continued for up to 7 days. Similar to Disperse trials, full inclusion and exclusion criteria for the MERLIN-TIMI36 trials (patient enrolment completed in May 2006), as well as study procedures, have been previously published [73].

Both the DISPERSE2-TIMI33 trial and the MERLIN-TIMI36 trial had long-term ECG data from patients used in our study. From these data, we measured the following ECG features: HRT categorized as low [turbulence onset (TO) <0 and turbulence slope (TS) >2.5 ms], moderate (either TO ≥ 0 or TS ≤ 2.5 ms), or high (TO ≥ 0 and TS ≤ 2.5 ms), DC [categorized as low (>4.5 ms), moderate (2.5 to 4.5 ms), or high (<2.5 ms)], and MV >50. In addition, the DISPERSE2-TIMI33 and the MERLIN-TIMI36 data sets had the following common clinical parameters: age >65 years, sex, current smoker, history of hypertension, history of diabetes mellitus, previous MI, index event (unstable angina versus nonST-elevation MI), ST-segment depression ≥ 0.1 mV.

A.2 BMC2

The Blue Cross Blue Shield of Michigan Cardiovascular Consortium (BMC2) multicenter interventional cardiology registry [74] collects data from all nonfederal hospitals that perform PCI in the state of Michigan. The BMC2 is a physician-run quality improvement collaborative that is supported by, but independent of, the funding agency, Blue Cross Blue Shield of Michigan. A physician advisory committee is responsible for setting the quality goals and developing quality improvement efforts without any input from or sharing of data with the study sponsor. Procedural data on all consecutive patients undergoing PCI at participating hospitals are collected by dedicated data abstractors using standardized data collection forms. All data elements have been prospectively defined, and the protocol is approved by the local institutional review board at each hospital. In addition to a random audit of 2% of all cases, medical records of all patients undergoing multiple procedures or coronary artery bypass grafting and of patients who died in the hospital are reviewed routinely to ensure data accuracy.

There were a total of 18,993 patients undergoing percutaneous coronary intervention in 2007, 22,023 patient undergoing percutaneous coronary intervention in 2008, and 20,289 patients undergoing percutaneous coronary intervention in 2009. The clinical variables available include those related to patient characteristics (gender, body mass index, age), cardiac status (priority, staged percutaneous coronary intervention, salvage, ad hoc percutaneous coronary intervention, stable angina, cardiac arrest, unstable angina, high-risk noncardiac surgery, atypical angina, patient turned down for coronary artery bypass graft by surgeon), percutaneous coronary intervention in the setting of MI (primary percutaneous coronary intervention; symptom to percutaneous coronary intervention time: 0 to 6, 6 to 12, 12 to 24, and >24 hours of symptoms; percutaneous coronary intervention of infarct-related vessel; cardiogenic shock; recurrent ventricular tachycardia or ventricular fibrillation; post-infarct angina; lytic therapy), comorbidities (current smoker, hypertension, insulin-dependent diabetes, noninsulin-dependent diabetes, congestive heart failure, peripheral vascular disease, renal failure requiring dialysis, significant valve disease, current or recent gastrointestinal bleed, chronic obstructive pulmonary disorder, cerebrovascular disease, atrial fibrillation, history of cardiac arrest, previous MI, previous percutaneous coronary intervention), pre-procedure laboratory results (creatinine, hemoglobin), contraindications (aspirin, angiotensin-converting enzyme inhibitors, beta-blockers, cholesterol-lowering agents, clopidogrel), pre-procedure therapy (aspirin, intravenous heparin, lowmolecular weight heparin, bivalirudin, angiotensin-converting enzyme inhibitors, beta-blockers, calcium channel blockers, diuretics, coumadin, clopidogrel, thienopyridine, intra-aortic balloon pump, intubation), and cardiac anatomy and function (left main artery stenosis, ejection fraction, number of diseased vessels, left ventricular end-diastolic pressure, graft lesion, grafts with ≥70% stenosis, ostial lesion, moderate to heavy calcification, thrombus, and chronic total occlusion).

A.3 UM-AFIB

The University of Michigan Atrial Fibrillation Study (UM-AFIB) is an ongoing study to collect data from patients undergoing CABG, aortic, or valvular surgery at the University of Michigan Hospital. Over the first 10 months of this project (December 2012 to October 2013) data from almost a thousand patients was collected. The data collection was carried out on the entire patient cohort in the Cardiothoracic intensive care unit (ICU) and the operating room (OR) comprising adult (age ≥18 years) patients.

Two sets of continuous ECG data were collected from each patient. The first was recorded during operation in the OR, and the other consists of at least 24 hours of continuous ECG data recorded in the ICU. Both recordings have 4-lead ECG available (Lead 1, 2, 3, and a generic V-lead that we refer to as Lead 4) and are sampled at 240 Hz with 16-bit quantization. ECG data acquisition was carried out from patients undergoing routine monitoring by using proprietary software developed at the University of Michigan to extract data from the GE Unity Network system used at the University of Michigan. Specifically, the OR data was recorded using GE Solar 9500 while the ICU data was recorded using GE Solar 8000M.

In parallel, the patients were monitored for PAF, and each episode of PAF was confirmed by clinician review. This process of validating and logging PAF is already part of standard care at the University of Michigan. The ECG data and the observations of PAF were supplemented with detailed metadata for all patients, corresponding to the risk metrics that have been described in the literature for PAF. These include:

- **preoperative variables**: age, gender, previous atrial fibrillation history, hypertension, stenosis of right coronary artery, chronic obstructive pulmonary disease, use of digoxin, beta-blocking withdrawal effect, previous myocardial infarction, smoking, diabetes mellitus, left ventricular ejection fraction, and left atrial dimension,

- **intraoperative variables**: (where applicable) aortic cross-clamp time, graft number, graft position, and choice of alternative minimal invasive off-pump CABG versus conventional on-pump CABG, and

- **postoperative variables**: mechanical ventilation time, postoperative pneumonia, assistance of intra-aortic balloon pump, and total amount of administered fluids.

All of the collected data are de-identified and anonymized.

APPENDIX B

Clinical Utility

B.1 Hazard Ratio

In survival analysis, a hazard is the rate at which events happen, so that the probability of an event happening in a short time interval is the length of time multiplied by the hazard. Although the hazard function may vary with time, the assumption is that the predictor variables does not change with time. This is also called proportional hazard models for survival analysis which assumes that the hazard in one group is a constant proportion of the hazard in the other group. The proportion of these corresponds to the hazard ratio.

Specifically, Cox proportional hazards regression [61] is a popular semiparametric method which we applied to quantify the effect of predictor variables. The Cox proportional hazards model relates the hazard rate for individuals or items at the value X, to the hazard rate for individuals or items at the baseline value. The impact of the predictor variables is a loglinear regression. For a baseline relative to 0, this model corresponds to

$$\log(HR) = \log(\frac{h_X(t)}{h_0(t)}) = \sum_{i=1}^{p}(X_i * \beta_i) \tag{B.1}$$

where $h_X(t)$ is the hazard rate at X and $h_0(t)$ is the baseline hazard rate function.

The hazard ratio represents the relative risk of instant failure for individuals or items having the predictive variable value X compared to the ones having the baseline values. For example, if the predictive variable is smoking status, where nonsmoking is the baseline category, the hazard ratio shows the relative instant failure rate of smokers compared to the baseline category, that is, nonsmokers. The estimated hazard ratio for the effect of each explanatory variable is $exp(\beta)$, given all other variables are held constant, where β is the coefficient estimate for that variable.

B.2 Area Under ROC Curve (AUROC)

The receiver operating characteristic (ROC) curve is a graphical plot which illustrates the performance of a binary classifier system as its discrimination threshold is varied. It is created by plotting the fraction of true positives out of the total actual positives (TPR = true positive rate) vs. the fraction of false positives out of the total actual negatives (FPR = false positive rate), at various threshold settings. TPR is also known as sensitivity or recall in machine learning. The FPR is also known as the fall-out and can be calculated as one minus the more well known specificity. The ROC curve is then the sensitivity as a function of fall-out.

The AUROC [19] measures discrimination, that is, the ability of the test to correctly classify those with and without the disease. Consider the situation in which patients are already correctly classified into two groups. You randomly pick one from the disease group and one from the no-disease group and do the test on both. The patient with the more abnormal test result should be the one from the disease group. The area under the curve is the percentage of randomly drawn pairs for which this is true (that is, the test correctly classifies the two patients in the random pair).

Under clinical settings, the AUROC can reflect the ability of the different approaches to discriminate between patients who died during follow up and those that remained event free. It is widely used in medicine, and is generally considered the standard for evaluating risk stratification methods [3].

B.3 Net Reclassification Improvement and Integrated Discrimination Improvement

Pencina et al. [83] proposed two new ways of assessing improvement in model performance (i.e., as a replacement to looking at the difference between AUROC values).

The net reclassification improvement (NRI) focuses on reclassification tables constructed separately for participants with and without events, and quantifies the correct movement in categories: upwards for events and downwards for non-events. For events, we assign 1 for upward reclassification, -1 for downward. The opposite is done for non-events. We then sum the individual scores and divide by the number of individuals in each group. Denoting D as the event indicator, we define the NRI as:

$$NRI = [P(up|D = 1) - P(down|D = 1)] - [P(up|D = 0) - P(down|D = 0)] \quad \text{(B.2)}$$

The integrated discrimination improvement (IDI) does not require categories, and focuses on the differences between sensitivities (IS) and 'one minus specificities (IP)' (the two axes of ROC plot) for two models.

$$IDI = (IS_{new} - IS_{old}) - (IP_{new} - IP_{old}) \quad \text{(B.3)}$$

APPENDIX C

Publications

1. **Chih-Chun Chia**, and Zeeshan Syed, ”Using Adaptive Downsampling to Compare Time Series with Warping.” *IEEE International Conference on Data Mining Workshops (ICDMW)*, 2010.

2. **Chih-Chun Chia**, and Zeeshan Syed, ”Computationally Generated Cardiac Biomarkers: Heart Rate Patterns to Predict Death Following Coronary Attacks.” *SIAM International Conference on Data Mining (SDM)*, 2011.

3. **Chih-Chun Chia**, Ilan Rubinfeld, Benjamin M. Scirica, Sean McMillan, Hitinder S. Gurm, and Zeeshan Syed, ”Looking beyond historical patient outcomes to improve clinical models.” *Science Translational Medicine (STM)*, 2012

4. **Chih-Chun Chia**, Zahi Karam, Gyemin Lee, Ilan Rubinfeld, and Zeeshan Syed, ”Improving surgical models through one/two class learning,” *Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2012

5. **Chih-Chun Chia**, James Blum, Zahi Karam, Satinder Singh, and Zeeshan Syed, ”Predicting Postoperative Atrial Fibrillation from Independent ECG Com-

ponents," *Association for the Advancement of Artificial Intelligence Conference (AAAI)*, 2014

6. **Chih-Chun Chia**, and Zeeshan Syed, "Scalable Noise Mining in Long-Term Electrocardiographic Time-Series to Predict Death Following Heart Attacks", *International conference on Knowledge Discovery and Data Mining (SIGKDD)*, 2014.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] R. Agrawal, C. Faloutsos, and A. N. Swami. Efficient similarity search in sequence databases. In *Proceedings of the 4th International Conference on Foundations of Data Organization and Algorithms*, FODO '93, pages 69–84, London, UK, UK, 1993. Springer-Verlag.

[2] G. Almassi, T. Schowalter, A. Nicolosi, A. Aggarwal, T. Moritz, W. Henderson, R. Tarazi, A. Shroyer, G. Sethi, F. Grover, et al. Atrial fibrillation after cardiac surgery: a major morbid event? *Annals of surgery*, 226(4):501, 1997.

[3] D. Altman. *Practical statistics for medical research*. Chapman & Hall/CRC, 1991.

[4] A. Andoni and P. Indyk. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. In *Foundations of Computer Science, 2006. FOCS'06. 47th Annual IEEE Symposium on*, pages 459–468. IEEE, 2006.

[5] G. Andrikopoulos, P. Dilaveris, D. Richter, E. Gialafos, A. Synetos, and J. Gialafos. Increased variance of p wave duration on the electrocardiogram distinguishes patients with idiopathic paroxysmal atrial fibrillation. *Pacing and Clinical Electrophysiology*, 23(7):1127–1132, 2006.

[6] S. Aranki, D. Shaw, D. Adams, R. Rizzo, G. Couper, M. VanderVliet, J. Collins, L. Cohn, and H. Burstin. Predictors of atrial fibrillation after coronary artery surgery: current trends and impact on hospital resources. *Circulation*, 94(3):390–397, 1996.

[7] C. Asher, D. Miller, R. Grimm, D. Cosgrove 3rd, M. Chung, et al. Analysis of risk factors for development of atrial fibrillation early after cardiac valvular surgery. *The American journal of cardiology*, 82(7):892, 1998.

[8] K. Aytemir, S. Aksoyek, N. Ozer, S. Aslamaci, and A. Oto. Atrial fibrillation after coronary artery bypass surgery: P wave signal averaged ecg, clinical and angiographic variables in risk assessment. *International journal of cardiology*, 69(1):49–56, 1999.

[9] K. Aytemir, N. ÖZER, E. Atalar, E. Sade, S. AKSÖYEK, K. ÖVÜNÇ, A. Oto, F. ÖZMEN, and S. Kes. P wave dispersion on 12-lead electrocardiography in patients with paroxysmal atrial fibrillation. *Pacing and Clinical Electrophysiology*, 23(7):1109–1112, 2000.

[10] T. L. Bailey and C. Elkan. The value of prior knowledge in discovering motifs with meme. In *Ismb*, volume 3, pages 21–29, 1995.

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

[12] A. Bauer, P. Barthel, R. Schneider, K. Ulm, A. Müller, A. Joeinig, R. Stich, A. Kiviniemi, K. Hnatkova, H. Huikuri, et al. Improved stratification of autonomic regulation for risk prediction in post-infarction patients with preserved left ventricular function (isar-risk). *European heart journal*, 30(5):576–583, 2009.

[13] A. Bauer, J. Kantelhardt, P. Barthel, R. Schneider, T. Mäkikallio, K. Ulm, K. Hnatkova, A. Schömig, 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.

[14] S. Ben-Haim, B. Becker, Y. Edoute, M. Kochanovski, and O. Azaria. Beat-to-beat electrocardiographic morphology variation in healed myocardial infarction. *The American journal of cardiology*, 68(8):725–728, 1991.

[15] C. Bennett. A linear programming approach to novelty detection. *Advances in Neural Information Processing Systems 13*, 13:395, 2001.

[16] D. Berndt and J. Clifford. Using dynamic time warping to find patterns in time series. In *AAAI94 workshop on knowledge discovery in databases*, pages 359–370, 1994.

[17] F. BJ. Defibrillators are lifesaver, but risks give pause. *New York Times*, 2008.

[18] D. Bloomfield, R. Steinman, P. Namerow, M. Parides, J. Davidenko, E. Kaufman, T. Shinn, A. Curtis, J. Fontaine, D. Holmes, et al. Microvolt t-wave alternans distinguishes between patients likely and patients not likely to benefit from implanted cardiac defibrillator therapy. *Circulation*, 110(14):1885–1889, 2004.

[19] A. P. Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997.

[20] A. Buxton and M. Josephson. The role of p wave duration as a predictor of postoperative atrial arrhythmias. *CHEST Journal*, 80(1):68–73, 1981.

[21] C. Cannon, S. Husted, R. Harrington, B. Scirica, H. Emanuelsson, G. Peters, R. Storey, et al. 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.

[22] P. Caravelli, M. Carlo, G. Musumeci, G. Tartarini, G. Gherarducci, U. Bortolotti, M. Mariani, et al. P-wave signal-averaged electrocardiogram predicts atrial fibrillation after coronary artery bypass grafting. *Annals of noninvasive electrocardiology*, 7(3):198–203, 2006.

[23] F. Castells, J. Igual, J. Rieta, C. Sanchez, and J. Millet. Atrial fibrillation analysis based on ica including statistical and temporal source information. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, volume 5, pages V–93. IEEE, 2003.

[24] C. Chang, S. Lee, M. Lu, C. Lin, H. Chao, J. Cheng, K. PEILIANG, and C. Hung. The role of p wave in prediction of atrial fibrillation after coronary artery surgery. *International journal of cardiology*, 68(3):303–308, 1999.

[25] N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer. Smote: synthetic minority over-sampling technique. *arXiv preprint arXiv:1106.1813*, 2011.

[26] C.-C. Chia and Z. Syed. Using adaptive downsampling to compare time series with warping. In *Data Mining Workshops (ICDMW), 2010 IEEE International Conference on*, pages 1304–1311. IEEE, 2010.

[27] G. Clifford, F. Azuaje, and P. McSharry. Advanced methods and tools for ecg data analysis. *Cambridge, Massachusetts*, 2006.

[28] M. Datar, N. Immorlica, P. Indyk, and V. Mirrokni. Locality-sensitive hashing scheme based on p-stable distributions. In *Proceedings of the twentieth annual symposium on Computational geometry*, pages 253–262. ACM, 2004.

[29] E. DeLong, D. DeLong, and D. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, pages 837–845, 1988.

[30] P. Dilaveris, E. Gialafos, S. Sideris, A. Theopistou, G. Andrikopoulos, M. Kyriakidis, J. Gialafos, and P. Toutouzas. Simple electrocardiographic markers for the prediction of paroxysmal idiopathic atrial fibrillation. *American heart journal*, 135(5):733–738, 1998.

[31] C. Dimmer, R. Tavernier, N. Gjorgov, G. Van Nooten, D. Clement, and L. Jordaens. Variations of autonomic tone preceding onset of atrial fibrillation after coronary artery bypass grafting. *The American journal of cardiology*, 82(1):22–25, 1998.

[32] N. El-Sherif, R. Mehra, W. Gough, and R. Zeiler. Reentrant ventricular arrhythmias in the late myocardial infarction period. interruption of reentrant circuits by cryothermal techniques. *Circulation*, 68(3):644–656, 1983.

[33] C. Elkan. The foundations of cost-sensitive learning. In *International Joint Conference on Artificial Intelligence*, volume 17, pages 973–978. Citeseer, 2001.

[34] T. Evgeniou and M. Pontil. Regularized multi–task learning. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 109–117. ACM, 2004.

[35] M. Fukunami, T. Yamada, M. Ohmori, K. Kumagai, K. Umemoto, A. Sakai, N. Kondoh, T. Minamino, and N. Hoki. Detection of patients at risk for paroxysmal atrial fibrillation during sinus rhythm by p wave-triggered signal-averaged electrocardiogram. *Circulation*, 83(1):162–169, 1991.

[36] C. Furberg, B. Psaty, T. Manolio, J. Gardin, V. Smith, and P. Rautaharju. Prevalence of atrial fibrillation in elderly subjects (the cardiovascular health study). *The American journal of cardiology*, 74(3):236–241, 1994.

[37] J. Glancy, C. Garratt, K. Woods, D. De Bono, et al. Qt dispersion and mortality after myocardial infarction. *Lancet*, 345(8955):945, 1995.

[38] J. Goldberger, M. Cain, S. Hohnloser, A. Kadish, B. Knight, M. Lauer, B. Maron, R. Page, R. Passman, D. Siscovick, et al. American heart association/american college of cardiology foundation/heart rhythm society scientific statement on noninvasive risk stratification techniques for identifying patients at risk for sudden cardiac death. *Circulation*, 118(14):1497–1518, 2008.

[39] S. Gottlieb, A. Dudek, D. Lowry, S. Nolan, and T. Guarnieri. Intravenous amiodarone for the prevention of atrial fibrillation after open heart surgery: the amiodarone reduction in coronary heart (arch) trial. *Journal of the American College of Cardiology*, 34(2):343–347, 1999.

[40] T. Guarnieri. Intravenous antiarrhythmic regimens with focus on amiodarone for prophylaxis of atrial fibrillation after open heart surgery. *The American journal of cardiology*, 84(9):152–155, 1999.

[41] P. Hamilton and W. Tompkins. Quantitative investigation of qrs detection rules using the mit/bih arrhythmia database. *IEEE transactions on bio-medical engineering*, 33(12):1157, 1986.

[42] M. Hauskrecht, M. Valko, B. Kveton, S. Visweswaran, and G. Cooper. Evidence-based anomaly detection in clinical domains. In *AMIA Annual Symposium Proceedings*, volume 2007, page 319. American Medical Informatics Association, 2007.

[43] H. He and E. Garcia. Learning from imbalanced data. *Knowledge and Data Engineering, IEEE Transactions on*, 21(9):1263–1284, 2009.

[44] T. Hiraki, H. Ikeda, M. Ohga, T. KUBARA, T. YOSHIDA, H. AJISAKA, A. TANABE, M. KANAHARA, and T. IMAIZUMI. Frequency-and time-domain analysis of p wave in patients with paroxysmal atrial fibrillation. *Pacing and clinical electrophysiology*, 21(1):56–64, 1998.

[45] C. Hogue Jr, P. Domitrovich, P. Stein, G. Despotis, L. Re, R. Schuessler, R. Kleiger, and J. Rottman. Rr interval dynamics before atrial fibrillation in patients after coronary artery bypass graft surgery. *Circulation*, 98(5):429–434, 1998.

[46] C. Hogue Jr, M. Hyder, et al. Atrial fibrillation after cardiac operation: risks, mechanisms, and treatment. *The Annals of thoracic surgery*, 69(1):300, 2000.

[47] K. Jongnarangsin and H. Oral. Postoperative atrial fibrillation. *Cardiology clinics*, 27(1):69–78, 2009.

[48] M. Josephson. *Clinical cardiac electrophysiology: techniques and interpretations*. Lippincott Williams & Wilkins, 2008.

[49] E. Keogh, K. Chakrabarti, M. Pazzani, and S. Mehrotra. Dimensionality reduction for fast similarity search in large time series databases. *Knowledge and information Systems*, 3(3):263–286, 2001.

[50] E. Keogh and C. A. Ratanamahatana. Exact indexing of dynamic time warping. *Knowledge and information systems*, 7(3):358–386, 2005.

[51] E. J. Keogh and M. J. Pazzani. Scaling up dynamic time warping for datamining applications. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 285–289. ACM, 2000.

[52] S. Khuri. The nsqip: a new frontier in surgery. *Surgery*, 138(5):837–843, 2005.

[53] M. Klein, S. Evans, S. Blumberg, L. Cataldo, and M. Bodenheimer. Use of p-wave-triggered, p-wave signal-averaged electrocardiogram to predict atrial fibrillation after coronary artery bypass surgery. *American heart journal*, 129(5):895–901, 1995.

[54] I. Kononenko. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in medicine*, 23(1):89–109, 2001.

[55] F. Korn, H. V. Jagadish, and C. Faloutsos. Efficiently supporting ad hoc queries in large datasets of time sequences. *ACM SIGMOD Record*, 26(2):289–300, 1997.

[56] H. Krumholz, P. Douglas, L. Goldman, and C. Waksmonski. Clinical utility of transthoracic two-dimensional and doppler echocardiography. *Journal of the American College of Cardiology*, 24(1):125–131, 1994.

[57] M. Kuhn, H. Tomaschewski, and H. Ney. Fast nonlinear time alignment for isolated word recognition. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'81.*, volume 6, pages 736–740. IEEE, 1981.

[58] J. Laurikkala, M. Juhola, E. Kentala, N. Lavrac, S. Miksch, and B. Kavsek. Informal identification of outliers in medical data. In *Proceedings of the 5th International Workshop on Intelligent Data Analysis in Medicine and Pharmacology*, pages 20–24. Citeseer, 2000.

[59] L. Lilly. *Pathophysiology of heart disease.* Lippincott, Williams & Wilkins, 2003.

[60] L. Lilly. *Pathophysiology of Heart Disease:: A Collaborative Project of Medical Students and Faculty.* Lippincott Williams & Wilkins, 2010.

[61] D. Y. Lin and L.-J. Wei. The robust inference for the cox proportional hazards model. *Journal of the American Statistical Association*, 84(408):1074–1078, 1989.

[62] J. Lin, E. Keogh, S. Lonardi, and B. Chiu. A symbolic representation of time series, with implications for streaming algorithms. In *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, pages 2–11. ACM, 2003.

[63] D. Lloyd-Jones, R. Adams, T. Brown, M. Carnethon, S. Dai, G. De Simone, T. Ferguson, E. Ford, K. Furie, C. Gillespie, et al. Heart disease and stroke statistics - 2010 update. *Circulation*, 121(7):e46–e215, 2010.

[64] N. Lomb. Least-squares frequency analysis of unequally spaced data. *Astrophysics and space science*, 39(2):447–462, 1976.

[65] A. L. W. M. E. Josephson. Fractionated electrical activity and continuous electrical activity: fact or artifact? *Circulation*, 70(3):529–32, 1984.

[66] W. Maisel, J. Rawn, and W. Stevenson. Atrial fibrillation after cardiac surgery. *Transplantation*, 151136:11, 2001.

[67] 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.

[68] M. Malik and A. Camm. *Dynamic electrocardiography.* Wiley Online Library, 2004.

[69] J. Mathew, M. Fontes, I. Tudor, J. Ramsay, P. Duke, C. Mazer, P. Barash, P. Hsu, D. Mangano, et al. A multicenter risk index for atrial fibrillation after cardiac surgery. *JAMA: the journal of the American Medical Association*, 291(14):1720–1729, 2004.

[70] J. Mathew, R. Parks, J. Savino, A. Friedman, C. Koch, D. Mangano, and W. Browner. Atrial fibrillation following coronary artery bypass graft surgery. *JAMA: the journal of the American Medical Association*, 276(4):300–306, 1996.

[71] A. Mehta, A. Jain, M. Mehta, and M. Billie. Usefulness of left atrial abnormality for predicting left ventricular hypertrophy in the presence of left bundle branch block. *The American journal of cardiology*, 85(3):354–359, 2000.

[72] L. Mitchell, D. Exner, D. Wyse, C. Connolly, G. Prystai, A. Bayes, W. Kidd, T. Kieser, J. Burgess, A. Ferland, et al. Prophylactic oral amiodarone for the prevention of arrhythmias that begin early after revascularization, valve replacement, or repair. *JAMA: the journal of the American Medical Association*, 294(24):3093–3100, 2005.

[73] D. Morrow, B. Scirica, E. Karwatowska-Prokopczuk, S. Murphy, A. Budaj, S. Varshavsky, A. Wolff, A. Skene, C. McCabe, and E. Braunwald. Effects of ranolazine on recurrent cardiovascular events in patients with non–st-elevation acute coronary syndromes. *JAMA: the journal of the American Medical Association*, 297(16):1775, 2007.

[74] M. Moscucci, D. Share, E. KLINE-ROGERS, M. O'DONNELL, A. MAXWELL-EWARD, W. L. Meengs, V. L. Clark, P. Kraft, A. C. FRANCO, J. L. Chambers, et al. The blue cross blue shield of michigan cardiovascular consortium (bmc2) collaborative quality improvement initiative in percutaneous coronary interventions. *Journal of interventional cardiology*, 15(5):381–386, 2002.

[75] R. Myerburg. Sudden cardiac death: exploring the limits of our knowledge. *Journal of Cardiovascular Electrophysiology*, 12(3):369–381, 2001.

[76] C. Myers, L. Rabiner, and A. Rosenberg. Performance tradeoffs in dynamic time warping algorithms for isolated word recognition. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 28(6):623–635, 1980.

[77] A. Naït-Ali. *Advanced biosignal processing.* Springer, 2009.

[78] K. Nigam, A. McCallum, S. Thrun, and T. Mitchell. Text classification from labeled and unlabeled documents using em. *Machine learning*, 39(2):103–134, 2000.

[79] T. F. of the European Society of Cardiology, the North American Society of Pacing, and Electrophysiology. Heart rate variability: standards of measurement, physiological interpretation, and clinical use. *Circulation*, 93:1043–1065, 1996.

[80] S. Ommen, J. Odell, and M. Stanton. Atrial arrhythmias after cardiothoracic surgery. *New England Journal of Medicine*, 336(20):1429–1434, 1997.

[81] S. Pan and Q. Yang. A survey on transfer learning. *Knowledge and Data Engineering, IEEE Transactions on*, 22(10):1345–1359, 2010.

[82] R. Passman, J. Beshai, B. Pavri, and S. Kimmel. Predicting post–coronary bypass surgery atrial arrhythmias from the preoperative electrocardiogram. *American heart journal*, 142(5):806–810, 2001.

[83] M. Pencina, R. D'Agostino Sr, R. D'Agostino Jr, and R. Vasan. Evaluating the added predictive ability of a new marker: from area under the roc curve to reclassification and beyond. *Statistics in medicine*, 27(2):157–172, 2008.

[84] M. Podgoreanu and J. Mathew. Prophylaxis against postoperative atrial fibrillation. *JAMA: the journal of the American Medical Association*, 294(24):3140–3142, 2005.

[85] K. Reddy. Cardiovascular disease in non-western countries. *New England Journal of Medicine*, 350(24):2438–2440, 2004.

[86] S. Roberts and L. Tarassenko. A probabilistic resource allocating network for novelty detection. *Neural Computation*, 6(2):270–284, 1994.

[87] H. Sakoe and S. Chiba. Readings in speech recognition. *Chapter Dynamic programming algorithm optimization for spoken word recognition*, pages 159–165, 1990.

[88] S. Salvador and P. Chan. Toward accurate dynamic time warping in linear time and space. *Intelligent Data Analysis*, 11(5):561–580, 2007.

[89] P. Schilling, J. Dimick, and J. Birkmeyer. Prioritizing quality improvement in general surgery. *Journal of the American College of Surgeons*, 207(5):698–704, 2008.

[90] A. SCHOENENBERGER, P. Erne, S. Ammann, G. Gillmann, R. Kobza, and A. STUCK. Prediction of arrhythmic events after myocardial infarction based on signal-averaged electrocardiogram and ejection fraction. *Pacing and clinical electrophysiology*, 31(2):221–228, 2008.

[91] B. Scholkopf, C. Burges, and A. Smola. Advances in kernel methods: Support vector machines, 1998.

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

[93] M. Shlipak, J. Ix, K. Bibbins-Domingo, F. Lin, and M. Whooley. Biomarkers to predict recurrent cardiovascular disease: the heart and soul study. *The American journal of medicine*, 121(1):50–57, 2008.

[94] S. Sovilj, A. Van Oosterom, G. Rajsman, and R. Magjarevic. Ecg-based prediction of atrial fibrillation development following coronary artery bypass grafting. *Physiological measurement*, 31(5):663, 2010.

[95] P. Stafford, J. Cooper, J. Fothergill, F. Schlindwein, C. Garratt, et al. Reproducibility of the signal averaged p wave: time and frequency domain analysis. *Heart*, 77(5):412–416, 1997.

[96] P. Stafford, P. Denbigh, and R. Vincent. Frequency analysis of the p wave: comparative techniques. *Pacing and Clinical Electrophysiology*, 18(2):261–270, 1995.

[97] P. Stafford, I. Turner, and R. Vincent. Quantitative analysis of signal-averaged p waves in idiopathic paroxysmal atrial fibrillation. *The American journal of cardiology*, 68(8):751–755, 1991.

[98] J. Steinberg, S. Zelenkofske, S. Wong, M. Gelernt, R. Sciacca, and E. Menchavez. Value of the p-wave signal-averaged ecg for predicting atrial fibrillation after cardiac surgery. *Circulation*, 88(6):2618–2622, 1993.

[99] G. D. Stormo and G. W. Hartzell. Identifying protein-binding sites from unaligned dna fragments. *Proceedings of the National Academy of Sciences*, 86(4):1183–1187, 1989.

[100] Z. Syed and J. Guttag. Identifying patients at risk of major adverse cardiovascular events using symbolic mismatch. *Ann Arbor*, 1001:48109, 2010.

[101] 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 Applied Signal Processing*, 2007(1):97–97, 2007.

[102] Z. Syed, P. Indyk, and J. Guttag. Learning approximate sequential patterns for classification. *The Journal of Machine Learning Research*, 10:1913–1936, 2009.

[103] Z. Syed, B. Scirica, S. Mohanavelu, P. Sung, E. Michelson, C. Cannon, P. Stone, C. Stultz, and J. Guttag. Relation of death within 90 days of non-st-elevation acute coronary syndromes to variability in electrocardiographic morphology. *The American Journal of Cardiology*, 103(3):307–311, 2009.

[104] Z. Syed, B. Scirica, C. Stultz, and J. Guttag. Risk-stratification following acute coronchia2010usingary syndromes using a novel electrocardiographic technique to measure variability in morphology. In *Computers in Cardiology, 2008*, pages 13–16. IEEE, 2008.

[105] 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 (TKDD)*, 4(1):2, 2010.

[106] Z. Syed, P. Sung, B. Scirica, D. Morrow, C. Stultz, and J. Guttag. Spectral energy of ecg morphologic differences to predict death. *Cardiovascular Engineering*, 9(1):18–26, 2009.

[107] L. Tarassenko, P. Hayton, N. Cerneaz, and M. Brady. Novelty detection for the identification of masses in mammograms. In *Artificial Neural Networks, 1995., Fourth International Conference on*, pages 442–447. IET, 1995.

[108] G. Thijs, K. Marchal, M. Lescot, S. Rombauts, B. De Moor, P. Rouzé, and Y. Moreau. A gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. *Journal of Computational Biology*, 9(2):447–464, 2002.

[109] V. Vapnik. Statistical learning theory. 1998, 1998.

[110] V. Vassilikos, G. Dakos, I. Chouvarda, L. Karagounis, H. Karvounis, N. Maglaveras, S. Mochlas, P. Spanos, and G. Louridas. Can p wave wavelet analysis predict atrial fibrillation after coronary artery bypass grafting? *Pacing and clinical electrophysiology*, 26(1p2):305–309, 2003.

[111] N. Weissman, A. Katz, and Y. Zigel. A new method for atrial electrical activity analysis from surface ecg signals using an energy ratio measure. In *Computers in Cardiology, 2009*, pages 573–576. IEEE, 2009.

[112] L. Ye and E. Keogh. Time series shapelets: a new primitive for data mining. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 947–956. ACM, 2009.

[113] A. Zaman, F. Alamgir, T. Richens, R. Williams, M. Rothman, and P. Mills. The role of signal averaged p wave duration and serum magnesium as a combined predictor of atrial fibrillation after elective coronary artery bypass surgery. *Heart*, 77(6):527–531, 1997.

[114] A. Zaman, R. Archbold, G. Helft, E. Paul, N. Curzen, and P. Mills. Atrial fibrillation after coronary artery bypass surgery: a model for preoperative risk stratification. *Circulation*, 101(12):1403–1408, 2000.

[115] W. Zong, G. Moody, and D. Jiang. A robust open-source algorithm to detect onset and duration of qrs complexes. In *Computers in Cardiology, 2003*, pages 737–740. Ieee, 2003.