Computational Methods for Physiological Data

by

Zeeshan Hassan Syed

S.B., Massachusetts Institute of Technology (2003)
M.Eng., Massachusetts Institute of Technology (2003)

Submitted to the Harvard-MIT Division of Health Sciences and Technology, and the MIT Department of Electrical Engineering and Computer Science

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Computer Science and Biomedical Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2009

Author . . . .
Harvard-MIT Division of Health Sciences and Technology, and the
MIT Department of Electrical Engineering and Computer Science
July 31, 2009

Certified by . . . .
John V. Guttag
Professor, Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by . . . .
Ram Sasisekharan
Director, Harvard-MIT Division of Health Sciences and Technology

Computational Methods for Physiological Data

by

Zeeshan Hassan Syed

Submitted to the Harvard-MIT Division of Health Sciences and Technology, and the
MIT Department of Electrical Engineering and Computer Science
on July 31, 2009, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Computer Science and Biomedical Engineering

Abstract

Large volumes of continuous waveform data are now collected in hospitals. These datasets provide an opportunity to advance medical care, by capturing rare or subtle phenomena associated with specific medical conditions, and by providing fresh insights into disease dynamics over long time scales.

We describe how progress in medicine can be accelerated through the use of sophisticated computational methods for the structured analysis of large multi-patient, multi-signal datasets. We propose two new approaches, *morphologic variability* (MV) and *physiological symbolic analysis*, for the analysis of continuous long-term signals. MV studies subtle micro-level variations in the shape of physiological signals over long periods. These variations, which are often widely considered to be noise, can contain important information about the state of the underlying system. Symbolic analysis studies the macro-level information in signals by abstracting them into symbolic sequences. Converting continuous waveforms into symbolic sequences facilitates the development of efficient algorithms to discover high risk patterns and patients who are outliers in a population.

We apply our methods to the clinical challenge of identifying patients at high risk of cardiovascular mortality (almost 30% of all deaths worldwide each year). When evaluated on ECG data from over 4,500 patients, high MV was strongly associated with both cardiovascular death and sudden cardiac death. MV was a better predictor of these events than other ECG-based metrics. Furthermore, these results were independent of information in echocardiography, clinical characteristics, and biomarkers. Our symbolic analysis techniques also identified groups of patients exhibiting a varying risk of adverse outcomes. One group, with a particular set of symbolic characteristics, showed a 23 fold increased risk of death in the months following a mild heart attack, while another exhibited a 5 fold increased risk of future heart attacks.

Thesis Supervisor: John V. Guttag
Title: Professor, Electrical Engineering and Computer Science

Acknowledgments

I chose John Guttag as an academic advisor at the end of freshman year, an exercise that entailed randomly selecting from a list of faculty members I had never met. For a decision made with precious little information, it has proven to be a pivotal one. Working with John has been one of my most cherished and enduring associations. He has been a teacher, mentor, colleague, friend, and family over the last many years, and his devotion to his students has been – surreal. None of the work presented in this thesis would have been possible without his inspiration and technical intuition, and his positive influence on my life and personality extends beyond the realm of research. It has been an absolute delight to have worked with John, and a pleasure to anticipate learning so much more from him in the years to come.

I also owe a great debt of gratitude to Collin Stultz, whose contributions to the work presented in this thesis are immeasurable. His dynamism and incredible knowledge of medicine and engineering have been invaluable while working at the intersection of these disciplines, and his unstinted support for our research has been a great comfort over the years. In addition, Collin has been a source of guidance on matters both technical and non-technical, and his dedication to the cause of others is an example that will remain with me for a long time.

I am indebted to Benjamin Scirica, who has been an esteemed colleague and wonderful friend during the course of this project. All aspects of our work have benefited immensely from Ben's varied skills. He has given direction to our research with his clinical insights and experience, helped us gain access to data that was instrumental in the development and testing of new hypotheses, suggested sophisticated statistical techniques to evaluate our work rigorously. and been an excellent sounding board for new ideas. Throughout the course of this project, Ben has helped accelerate progress and motivated us with his endless enthusiasm. I also thank Ben for introducing us to Christopher Cannon, Peter Stone, David Morrow, Satishkumar Mohanavelu, Eugene Braunwald and the rest of the TIMI Study Group at the Brigham and Women's Hospital. I am grateful to them for sharing their data and expertise with us, and consider myself fortunate to have worked with some of the finest clinical minds of our time.

I also feel privileged to have worked with George Verghese, who continues to be a role model with his extensive knowledge spanning many disciplines and his humility. His feedback and questions have helped identify important areas where significant improvements were possible, and have been essential in helping me develop a more complete grasp of complex engineering principles.

I am grateful to Piotr Indyk for the substantial algorithmic improvements in our work, most notably in the area of symbolic analysis, which were possible with his help. Piotr has always been available to help on short notice, and it has been fascinating to watch him reason about complicated theoretical issues with great ease and clarity.

Manolis Kellis has also helped by sharing his expertise on motif discovery methods. His in-depth knowledge and input have helped shape much of our work on computational physiology.

This work has also benefited greatly from Gari Clifford and his expertise in the analysis of electrocardiographic data. Gari has contributed many important tools

that were used in our project and almost all aspects of this thesis bears some imprint of his ideas.

Thomas Heldt has motivated our work on the pathophysiology of morphologic variability, and his suggestions on experiments to validate our theory represent an exciting area of future research. In addition, Thomas has been a source of encouragement and during the course of this project, has grown to be a good friend.

I thank Roger Mark, George Moody, and the Physionet Group for contributing the Physionet database, which formed the starting point of our work. The work in this thesis also owes greatly to Barbara Liskov and the Programming Methodology Group at MIT. Their kindness in sharing computational resources was critical in helping us develop, refine and validate our work, and has allowed the scale of our project to expand significantly.

I would also like to thank James Glass, Hari Balakrishnan, Tommi Jaakkola, Samuel Madden, Michael Collins, Dina Kitabi, Srini Devadas, and Ronitt Rubinfeld for their many suggestions to improve various aspects of this work, and for their help in determining how best to communicate our findings to the rest of the scientific community. The presence of so many brilliant and friendly individuals is an aspect of MIT that continues to amaze even after all these years.

I am grateful to Adam Wolfberg for creating the opportunity for us to extend our work to fetal datasets. Adam has added an important dimension to our research, and helped evolve our vision of how computational methods can make a broad clinical impact. His enthusiasm while solving the hard problem of fetal risk stratification has been inspirational.

I am also indebted to Charles Vest, Danielle Guichard-Ashbrook, Humayun Bashir, Umer Shah, Salik Malik, and Mubasher Humayun. Their support and guidance enabled me to return to MIT for my PhD, and I credit them with this work.

I take great pride in having been one of the initial members of the Data Driven Medicine Group, which has been a fabulous home for me over the years, and allowed me to meet some exceptional individuals. I owe a great deal both personally and professionally to Dorothy Curtis, Sheila Marian, Donna Kaufmann, John Ankcorn, Eugene Shih, Godfrey Tan, Irene Fan, Jenna Wiens, Anima Singh, Al Kharbouch, Daniel Leeds, Jason Gift, Marcia Jung, Joyatee Sarker, and Jennifer Carlisle. I also thank all the members of the Network and Mobile Systems group, who have shared many experiences with me for the last many years: Srikanth Kandula, Sachin Katti, Magda Balazinska, Michel Goraczko, Allen Miu, Kyle Jamieson, Hariharan Rahul, Vladimir Bychkovsky, Jaeyeon Jung, Nick Feamster, Bodhi Priyantha, and David Andersen.

I mention Phil Sung separately, for his contributions to this thesis are immense. He has been a close colleague over the last two years, and his brilliance and work ethic have led to significant improvements in the way that morphologic variability is measured. It was a pleasure working with him, and I sincerely hope our paths will cross again in the future.

Ali Shoeb appears late in this list, but only because he represents the bridge between my professional and social lives. I have enjoyed his company since the start of freshman year, taken nearly all of my classes at MIT with him, followed in his example by joining the Harvard-MIT Health Sciences and Technology program, and shared an office with him for as long as I can recall. I have often marveled at the range and depth of Ali's knowledge, and his ability to explain things with pristine clarity. His contributions to this thesis are too many to recount, and his suggestions pervade all aspects of our work.

It would be impossible to mention all of my other friends at MIT, but I thank Ebad, Murtaza, Daanish, Rehan, Saad, Elena, Sarfraz, Faisal, Imran, and Adnan, in particular, for contributing ideas that have gone into improving the work presented in this thesis, and for making my academic career at MIT a pleasure.

Finally, I dedicate my efforts to my family, who have made many sacrifices to allow me to be where I am today. My parents have taught me to advance myself at every step of life, value integrity, work hard, and demonstrate a concern for others. I am eternally indebted for everything that they have empowered me to achieve. I also owe thanks to my wife, whose consistent support and encouragement during the later stages of my PhD have been an immensely valuable asset.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this thesis, we present novel ways to improve medical care by using sophisticated computational methods to analyze clinical data. A rough problem statement for our work is to develop automated tools for the prediction of future health and treatment efficacy. There are two aspects to this: discovering knowledge from large multi-signal datasets collected from a population of patients, and applying this knowledge to developing automated methods for improved prognosis and intervention. This research uses techniques drawn from machine learning, data mining, applied algorithms, and signal processing; as well as an understanding of the underlying biological systems.

1.1 Opportunities

Our work is motivated by the increasing amounts of continuous long-term primary data available for patients in hospital and ambulatory settings. With advances in recording and storage technology, we are now able to collect larger volumes of data than was previously possible. This increase in available data has taken place both at the level of individual patients (i.e., with more types of data now being captured over longer periods), and at the level of the population as a whole (i.e., more patients being monitored with ambulatory devices).

These large physiological datasets present an opportunity to advance medical care along different dimensions. Monitoring many different patients continuously over long periods (i.e., from hours to weeks) increases the likelihood of discovering previously unknown phenomena, and helps provide fresh insights into disease dynamics over long time scales. We can therefore use continuous long-term data to discover new medical knowledge. In addition to this, continuous monitoring also makes it possible to observe known but rare events that would otherwise be missed if patient health was assessed in a point-in-time manner (e.g., patients who have hypertensive episodes at night are routinely missed by blood pressure measurements in the doctor's office [2]). We can therefore use continuous long-term data to apply known medical knowledge better. Both the discovery of new knowledge and the ability to apply knowledge better allow patients to be matched to therapies that are most appropriate for their risk.

1.2 Challenges

Despite the opportunities provided by large amounts of continuous long-term data, the sheer volume of information is a serious challenge. Patients in an ICU setting, for example, often have continuous streams of data arising from telemetry monitors, pulse oximeters, Swan-Ganz catheters, and arterial blood gas lines to name just a few sources. Any process that requires humans to examine more than small amounts of data is infeasible and often highly error prone. To put things in perspective, the electrocardiographic (ECG) signals from a single patient admitted to a hospital following a heart attack would fill 8,000 pages. It is therefore not surprising that errors have been associated with "information overload" and that clinically relevant events are often missed [70, 71].

Unfortunately, existing software systems are also largely inadequate for studying large physiological datasets. First, existing software systems focus largely on detection and are restricted in their ability to do broad knowledge discovery and identify previously unrecognized activity. They do not exploit the opportunity provided by large physiological datasets to discover new medical knowledge. What work there has been has largely focused on the analysis of categorical data (e.g., health records) or nucleotide sequences, and has not addressed the challenge of studying large datasets comprising continuous time-series signals.

Secondly, existing software systems are limited in their ability to apply medical knowledge. In particular, the choice of which detectors to apply to continuous data is based on assumptions about which events are most likely to occur (e.g., the use of arrhythmia monitors following a heart attack). This limits their ability to detect even known events that were not considered probable ahead of time.

Another challenge while working with continuous long-term signals is that of efficiency. The process of collecting data sampled at high frequencies over long periods (i.e., days to weeks) for thousands of patients leads to very large datasets. Analyzing these datasets with complex methods is computationally intractable. This creates the need for methods that are both robust and efficient.

1.3 Proposed Solutions

We propose two broad areas of complementary research for studying information in large physiological datasets. Our methods address the goals of both discovering and applying medical knowledge, and are intended to promote personalized medicine through more accurate risk stratification and the choice of interventions that are consistent with each patient's individual risk. Our methods achieve computational efficiency through use of algorithmic improvements and abstraction to reduce data.

At an abstract level, we view interesting activity in physiological signals as variability at different scales.

Morphologic variability focuses on the micro-level changes in continuous signals, and quantifies subtle variability in signals over long periods. This is intended as a way to measure instability in the underlying physiological system.

Symbolic analysis, looks at the macro-level information in signals by abstracting them into symbolic sequences and studying the resulting textual representations of the time series signals for interesting higher-level constructs.

In this thesis, we show how both approaches can be used separately or together to obtain clinically significant results.

1.4 Clinical Application

While the computational methods we develop in this thesis are applicable to signals from many different physiological systems, and potentially to data from non-medical applications, we present and evaluate these tools in the context of cardiovascular disease. There are a variety of factors that led us to this decision.

Cardiovascular disease is a key clinical area. According to the statistics reported for 2005, there were 864,500 deaths in the US due to cardiovascular disease [72], corresponding to 35.5% of all deaths. For the same period, an estimated 17.5 million people died from cardiovascular disease around the world (30% of all global deaths) [3]. Nearly 151,000 individuals in the US who experienced cardiovascular mortality in 2005 were under age 65, well below the average life expectancy of 77.8 years [72]. The associated direct and indirect costs of dealing with cardiovascular disease was $475.3 billion [72].

One of the challenges of modern cardiovascular medicine is to supplement advances in therapy with similar advances in diagnostics. An example of the present divide between treatment and diagnostics is provided by implantable cardioverter defibrillators (ICDs) [73]. These devices are typically given to the patients believed to be at the greatest risk, and have great value in reducing deaths when serious arrhythmias occur. However, according to some estimates, 90% of the patients who received ICDs never end up using these devices [73]. This represents a major inefficiency and danger, since these devices are expensive and require a surgical procedure (with a 1% operative risk of death) for implantation. At the other end of the spectrum are the hundreds of deaths that take place each day of patients who were truly at high risk of fatal arrhythmias, but were not correctly categorized and did not receive an ICD. Our work attempts to remedy this kind of situation, by using sophisticated computational methods to analyze large amounts of cardiovascular data to identify high (or low) risk patients.

The majority of our work focuses on a specific kind of cardiovascular signal, i.e., the electrocardiogram (ECG). This follows from the availability of large datasets of ECG signals collected routinely during hospitalization and clinical trials. Access to these signals and detailed follow-up data for patients has helped us rigorously evaluate our methods. The choice of ECG signals is also motivated by the fact that these signals have been extensively studied in the past. It is thus possible to demonstrate how sophisticated computational methods can be used to yield novel information even from well-studied sources of data. Finally, ECG signals are often collected for many days by ambulatory monitors. This has allowed us to demonstrate the utility of studying data collected over long periods of time using the right tools.

1.5 Contributions

We briefly review some of the major contributions of our work. A more detailed discussion on the different contributions made by our work is deferred to subsequent parts of the thesis.

- **Concept of Morphologic Variability**: We propose the concept of morphologic variability, and introduce the idea that subtle variations in the shape of physiological signals, which are often widely considered to be noise, contain important information about the health of the underlying system.

- **Method to Measure Morphologic Variability**: We develop a method to measure morphologic variability, which addresses the challenge of quantifying subtle pathological variability in noisy signals with time-skew. Our algorithm uses a modified dynamic time-warping approach to compare variations in morphology between consecutive beats, and the Lomb-Scargle periodogram to identify a spectral signature for these variations that corresponds to high risk.

- **Clinical Study of Morphologic Variability**: We conducted a clinical study on data from over 4,500 patients, and show that morphologic variability has considerable predictive value in the setting of cardiovascular disease. For example, patients with high morphologic variability are at a 6-7 fold increased risk of death in the three months following a heart attack. Moreover, the information in morphologic variability is independent of other generally accepted risk variables (e.g., echocardiography and other electrocardiographic metrics) and morphologic variability is a better predictor of death than almost all of these variables. In particular, it has great value in identifying patients who are missed by echocardiography, which is widely used to identify a small high risk group of patients but may miss over two-thirds of all deaths [1].

- **Concept of Physiological Symbolic Analysis**: We propose the concept of physiological symbolic analysis, i.e., representing and searching through continuous physiological waveform signals as textual data rather than real-valued time-series. We develop a symbolic analysis framework that allows for physiological datasets to be studied in a manner analogously to nucleotide data.

- **Method to Symbolize Physiological Signals**: We present an efficient Max-Min clustering-based algorithm for symbolization, and demonstrate this transformation preserves useful clinical information while making the data easier to analyze. We show how different analyses on symbolic representations can be used to detect various kinds of interesting activity, e.g., searching for approximate repeated sequences finds ventricular bigeminy and trigeminy; searching for statistically overrepresented patterns reveals tachyarrhythmias; and locating high entropy periods detects atrial fibrillation. We also demonstrate how these methods can be used to find kinds of complex activity that often go unnoticed in clinical practice, e.g., atrial ectopic rhythms.

- **Method for Discovering Predictors of Acute Events in Symbolic Data**: We present novel methods that can be used to discover predictors of acute events in an automated manner, by searching for approximate symbolic patterns that occur more often preceding events than one would expect by chance alone. We approach pattern discovery as a significance and classification problem, and use the ideas of locality sensitive hashing (LSH), multi-level Gibbs sampling, and sequential statistics to make the search for interesting activity more efficient.

- **Clinical Study of Predictor Discovery**: We demonstrate the utility of our methods to discover predictors of acute events, both for detecting markers associated with long-term risk and for markers associated with imminent acute events (e.g., sudden cardiac death). In a small study of patients who experienced sudden cardiac death, our algorithms correctly predicted 70% of the deaths while classifying none of the normal individuals and only 8% of the patients with supraventricular arrhythmias as being at risk.

- **Method for Comparative Risk Stratification with Symbolic Data**: We develop an algorithm to compare the long-term symbolic dynamics of patients by measuring the probability-weighted mismatch of symbol prototypes across patients to assess similarity. We present a second, clustering-based algorithm that uses this similarity information to partition patients into groups with similar risk profiles.

- **Clinical Study of Comparative Risk Stratification with Symbolic Data**: We evaluated our comparative methods on cardiac data to partition patients with cardiovascular disease into groups, and found that different groups of patients exhibit a varying risk of adverse outcomes. One group, with a particular set of time-series characteristics, showed a 23 fold increased risk of death in the months following a mild heart attack, while another exhibited a 5 fold increased risk of future heart attacks. This potentially allows more fine-grained risk assessment of patients.

- **Concept of Visualizing Physiological Signals as Symbolic Sequences**: We propose the concept of visualizing large amounts of continuous monitoring data as a sequence of symbols rather than raw samples, and develop tools to visualize continuous long-term signals. Looking at physiological data as symbolic sequences provides many advantages over visualizing raw samples. It results in a large decrease in the number of data points that need to be visualized, makes it easier to see when changes occur, and makes the data more readily interpretable. Our software supports visualizing continuous long-term signals as symbols, while retaining information (in the form of prototypes and even the raw data) that is available to users interested in looking at waveforms.

- **Concept and Method of Creating Prototypical Signals Averaged in Amplitude and Time**: We develop the idea of creating prototypical representations of physiological activity. In contrast to conventional aggregation approaches, which average the amplitude of multiple observations, we propose creating a signal where the duration of each physiological waveform was also averaged in time. We also propose an algorithm to create prototypical signals from noisy, time-skewed observations, by relating time-warped information across observations and combining data hierarchically while preserving length characteristics. We demonstrate how these prototypes can be used for both data visualization, and for robust decision making.

1.6 Organization of Thesis

The remainder of this thesis is organized as follows. Chapter 2 presents background on the heart and the ECG signal. Chapter 3 introduces the concept of morphologic variability and evaluates its use to identify high risk cardiovascular patients. Chapter 4 presents and evaluates symbolic analysis of large datasets of physiological signals. Chapter 5 describes tools to compactly visualize interesting activity over long periods of continuous patient monitoring. Chapter 6 concludes with a summary and discussion of future work.

Chapter 2

Background

In this chapter, we review the clinical background for our work. We start with a discussion of the normal anatomy and function of the heart in Section 2.1. We focus, in particular, on aspects of cardiac function related to electrophysiology. This is followed by a presentation of the normal electrocardiogram (ECG) signal in Section 2.2. Cardiovascular pathophysiology is then reviewed in Sections 2.3 and 2.4, which describe acute coronary syndromes (ACS) and arrhythmias respectively. Finally, we present a summary of existing methods to identify high risk patients in Section 2.5.

This material provides context for our work. The methods introduced in this thesis have largely been validated on ECG data from post-ACS patients, some of whom experience fatal arrhythmias. Readers may therefore find a discussion of the electrocardiogram, acute coronary syndromes, and arrhythmias to be helpful. Those readers already knowledgeable about these topics may want to skip to Section 2.5.3, which provides a review of existing methods to identify high risk cardiac patients. We compare our work with these methods later in this thesis.

This chapter borrows heavily from the discussion of these subjects in [68, 69].

2.1 Cardiac Function

The heart has four separate compartments or chambers. The upper chambers on either side of the heart, which are called atria, receive and collect the blood coming to the heart. The right atrium receives blood from the inferior and superior vena cava and the left atrium is supplied by the pulmonary veins. The atria then deliver blood to the powerful lower chambers, called ventricles, which pump blood away from the heart through powerful, rhythmic contractions. Blood leaves the right ventricle through the pulmonary artery and similarly, the left ventricle is connected to the aorta. The first branches of the aorta are small arteries known as the coronary arteries. These supply blood to the heart itself.

The heart circulates blood via the coordinated action of its chambers. Each heartbeat can be divided into two main stages: systole and diastole. During systole, the atria first contract, pushing a small fraction of their volume of blood into the ventricles to fill them to maximum capacity. This is followed by the contraction of the ventricles, which pushes blood out of the heart and into the pulmonary artery and the aorta. Diastole takes place once systole is complete. During this period, both the atria and ventricles are relaxed, and continue to fill with blood till the next systole occurs.

At rest, the heart beats roughly about 70 times per minute, with each beat having a corresponding duration of approximately 800 ms. The heart rate and the duration of each beat vary significantly among individuals and may also have different values for the same person depending on the activity being performed. The periodic pumping action of the heart results in the unidirectional flow of blood through the human body, and is known as the cardiac cycle. This process is coordinated by the orderly propagation of electrical impulses throughout the *myocardium*, or heart muscle, which causes these cells to contract.

In the remainder of this section, we focus on the electrophysiology of the heart. The normal conduction system of the heart is pictured in Figure 2-1. A wave of depolarization (i..e, a temporary reversal of the cell membrane voltage) begins in the sinoatrial (SA) node, which contains *pacemaker cells* that spontaneously produce electrical impulses. From there, depolarization spreads throughout the atria, causing them to contract. The wave then reaches the atrioventricular (AV) node. This is the only connection between the conduction systems of the atria and the ventricles,

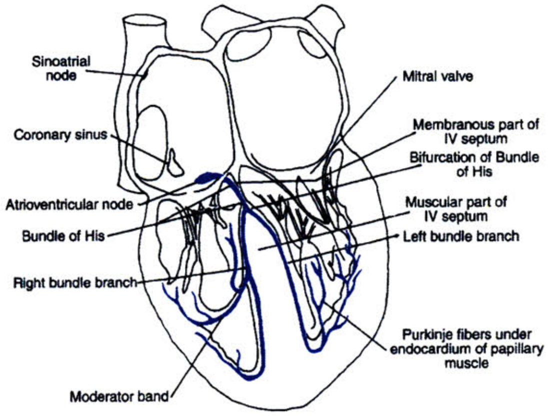

Sinoatrial node
Coronary sinus
Atrioventricular node
Bundle of His
Right bundle branch
Moderator band
Mitral valve
Membranous part of IV septum
Bifurcation of Bundle of His
Muscular part of IV septum
Left bundle branch
Purkinje fibers under endocardium of papillary muscle

Figure 2-1: Main components of the cardiac conduction system. From Lilly [95].

which are elsewhere separated by insulating fibrous tissue. The AV node consists of specialized tissue that conducts slowly, so it delays electrical impulses that pass through it for a short time (about 0.1 sec). This delay is important for efficient circulation because it allows the atria to completely empty their blood into the ventricles before the ventricles begin to contract. Finally, the wave of depolarization spreads throughout the ventricles by way of the Bundle of His and the left and right bundle branches, causing the ventricles to contract.

The electrical activity of the heart is associated with different changes at the cellular level. The membrane of a myocardial cell contains *ion channels*, specialized proteins that span the cell membrane and regulate the movement of specific ions across the membrane [95]. Different types of ion channels are selective for different kinds of ions, allowing only ions of a specific type to pass. In addition, the conformation of ion channels changes with the membrane voltage difference to allow (or block) the diffusion of ions. Ion channels act as voltage-regulated passive gates for ions: the flow of ions through ion channels is determined by the concentration gradient and by the electrical potential difference (voltage) across the membrane. Cell membranes also contain active *ion pumps*, which consume energy in the form of adenosine triphosphate (ATP) to pump ions across a membrane against their natural gradient.

In a cardiac cell at rest, the ion channels and ion pumps together maintain a *resting potential* of -90 mV inside the cell by selectively moving Na^+ and Ca^{++}

ions out of the cell and K^+ ions into the cell. If the membrane voltage goes above approximately -70 mV, an *action potential* begins. Some sodium ion channels open, allowing Na^+ ions to enter the cell, raising the potential inside, causing more sodium ion channels to open, and so on, creating a positive feedback loop. The cell quickly (within milliseconds) becomes depolarized and reaches a peak voltage of slightly more than 0 mV. This voltage is high enough to raise the membrane voltage in a nearby area of the cell or a neighboring cell, causing the action potential to propagate.

At the peak voltage, the sodium channels close and remain inactivated until the cell has returned to resting potential (as described below). In healthy myocardial tissue, this refractory period prevents recently depolarized cells from depolarizing again, regardless of the membrane voltage. This ensures that the wave of depolarization propagates forward and never backward.

The cell now begins the process of *repolarization* in order to prepare for the next action potential. When the membrane voltage becomes high enough, the potassium and calcium channels open, allowing K^+ and Ca^{++} ions to flow out of and into the cell, respectively. Calcium ions entering the cell during this phase activate a pathway that induces the physical contraction of cardiac muscle cells. Finally, the original concentrations of each ion, and the resting potential, are restored by ion pumps in order to prepare the cell for another action potential.

Abnormal cardiac depolarization and repolarization may lead to fatal arrhythmias, as will be discussed in Section 2.4. We believe that the techniques presented in this thesis can help detect problems with the electrical system of the heart, and quantify the extent of any present abnormalities.

2.2 Electrocardiogram

An electrocardiogram (ECG) is a recording of the electrical activity of the heart. ECG data is routinely recorded for hospitalized patients, since it is useful for both monitoring them and diagnosing conditions such as ACS or arrhythmias. ECG can be acquired inexpensively and with minimal invasiveness; a Holter monitor (a portable

ECG device worn on a patient) can record data for 24 hours or more. Therefore, ECG data is useful for analysis of rare and noisy phenomena. Depending on the setting and on the reason for the recording, varying numbers of electrodes may be used in order to capture a varying number of channels of data. Typical ECG monitors record between 1 and 12 channels.

A cardiac muscle cell at rest maintains a negative voltage with respect to the outside of the cell. While at rest, the surface of the cell is uniformly charged with a positive voltage, but during depolarization, this voltage decreases and may even become negative. Consequently, when depolarization is propagating through a cell, there exists a potential difference on the membrane between the part of the cell that has been depolarized and the part of the cell at resting potential. After the cell is completely depolarized, its membrane is uniformly charged again (although now negatively instead of positively).

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

Three major segments can be identified in a normal ECG, corresponding to different parts of the action potential. Schematics of the cardiac conduction pathway and a typical ECG recording are shown in Figure 2-2. The P wave is associated with depolarization of the atria. The QRS complex is associated with depolarization of the ventricles. The T wave is associated with repolarization of the ventricles. The QRS

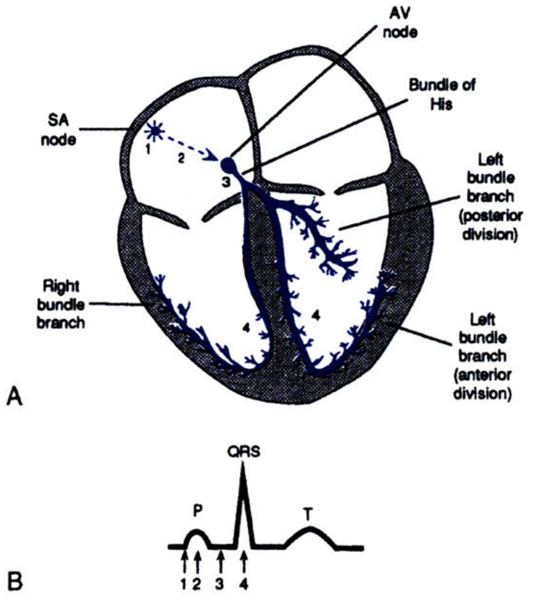

AV node
Bundle of His
SA node
1
2
3
Left bundle branch (posterior division)
Right bundle branch
4
4
Left bundle branch (anterior division)
A
QRS
P
T
1 2 3 4
B

Figure 2-2: Cardiac conduction pathway, with corresponding waveforms on the ECG recording. The electrical impulse begins at the SA node (1). The wave of depolarization traverses the atria (2). Conduction is slowed at the AV node (3). The wave of depolarization traverses the ventricles (4). From Lilly [95].

complex is larger than the P wave because the ventricles are much larger than the atria. The QRS complex coincides with 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.

2.3 Acute Coronary Syndromes

We follow the earlier discussion of normal cardiac function and electrophysiology with brief review of cardiac pathophysiology. In particular, we focus on acute coronary syndrome (ACS), an umbrella term covering clinical symptoms compatible with reduced blood supply to the heart (i.e., myocardial ischemia). Heart attacks and unstable angina are included in this group.

An acute coronary syndrome (ACS) is an event in which the blood supply to part of the myocardium is blocked or severely reduced. The most common symptom of ACS is unusual and unprovoked chest pain, but this may often be absent (most notably

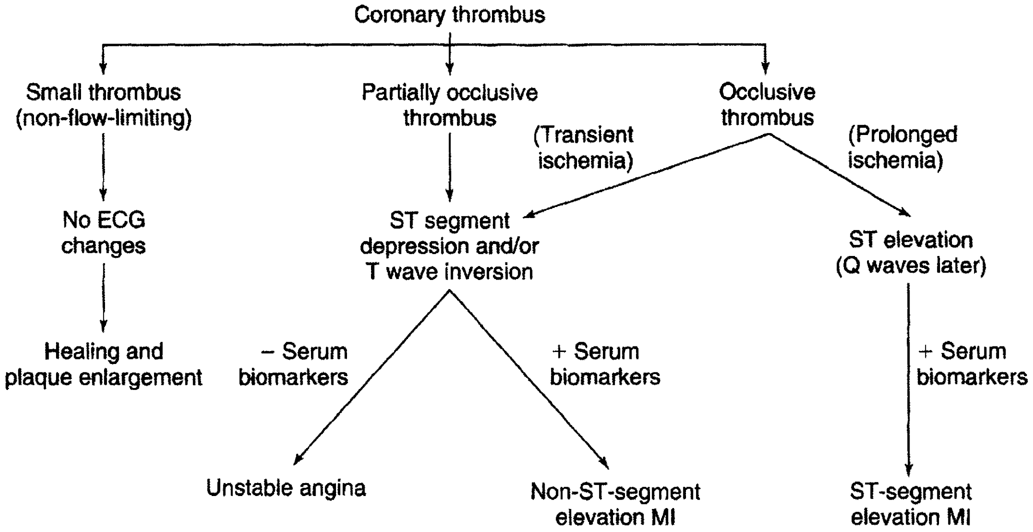

Coronary thrombus
Small thrombus (non-flow-limiting)
Partially occlusive thrombus
Occlusive thrombus
(Transient ischemia)
(Prolonged ischemia)
No ECG changes
ST segment depression and/or T wave inversion
ST elevation (Q waves later)
Healing and plaque enlargement
− Serum biomarkers
+ Serum biomarkers
+ Serum biomarkers
Unstable angina
Non-ST-segment elevation MI
ST-segment elevation MI

Figure 2-3: Consequences of coronary thrombosis. From Lilly [95].

in patients with diabetes who experience "silent" heart attacks). Other symptoms include shortness of breath, profuse sweating, and nausea.

An ACS is usually caused by the rupture of an atherosclerotic plaque producing a blood clot within a coronary artery. This restricts blood flow to the heart, causing ischemia and potentially cell death in the myocardium. Various subclassifications of ACS are distinguished by the presence of myocardial necrosis (cell death) and by ECG diagnosis. An overview of these subclassifications is shown in Figure 2-3.

Unstable angina refers to an ACS event in which necrosis does not occur, while myocardial infarction (MI) refers to one in which it does. An ACS is also subclassified based on the extent to which the coronary artery is occluded, which can often be inferred noninvasively from ECG recordings. An ECG showing elevation in the ST segment is indicative of complete occlusion of an artery and necrosis (and therefore, myocardial infarction). Such patients are given a diagnosis of ST-elevation MI (STEMI) and are typically higher risk relative to patients with non-ST-elevation ACS.

Non-ST-elevation ACS (NSTEACS) is indicative of partial occlusion of an artery and is a less severe condition. NSTEACS may be diagnosed by the presence of certain ECG irregularities (ST depression or T wave inversion). Two subclasses of

NSTEACS, unstable angina and a non-ST-elevation MI (NSTEMI), are distinguished by whether necrosis occurs. Blood tests are used to determine levels of two *serum biomarkers*, cardiac-specific troponin and creatine kinase MB (CK-MB), which are chemicals released into the bloodstream when myocardial necrosis occurs.

Treatment for NSTEACS focuses on inducing the dissolution of blood clots by natural pathways (via aspirin or heparin), and on reducing ischemia by lowering the heart's oxygen demand and raising oxygen supply. Drugs that dilate blood vessels (nitrates) or lower heart rate (β-blockers) are commonly employed. STEMI patients may benefit from the same treatments, but they also receive more aggressive thrombolytic drugs to break down blood clots and restore normal blood flow. Percutaneous coronary intervention (PCI) and coronary artery bypass graft (CABG) may also be conducted, either immediately or after the patient's condition has stabilized.

2.4 Arrhythmias

An ACS may leave damaged or scarred heart tissue, which can interfere with the heart's electrical conduction system. This may lead to arrhythmias, i.e., abnormal heart rhythms. The heart may beat too fast or too slowly, and may be regular or irregular.

Some arrhythmias are life-threatening medical emergencies that can result in cardiac arrest and sudden death. Others cause symptoms such as palpitations, while still others may not be associated with any symptoms at all, but predispose toward potentially life-threatening stroke or embolus. Arrhythmias are generally classified into two groups: tachyarrhythmias (where the heart beats too quickly) and bradyarrhythmias (where the heart beats too slowly). These may arise from irregularities in the generation of action potentials, or in the conduction of action potentials through the myocardium.

The generation of action potentials is usually the job of the SA node. In abnormal situations, other parts of the heart may start to spontaneously depolarize (leading to tachyarrhythmias) or impulse generation may be impaired (leading to bradyarrhythmias). Typically, a bradyarrhythmia stemming from impaired impulse generation is not a life-threatening situation, because the myocardium contains multiple regions of tissue that have the potential to spontaneously depolarize; these act as "backup" pacemakers if impulse generation at the SA node becomes too slow.

Major conduction pathway alterations can also lead to arrhythmias. A *conduction block* arises when a region of unexcitable tissue stops the wave of depolarization entirely, preventing part of the heart from contracting. *Reentry* is a phenomenon in which a wave of depolarization travels around a closed-loop conduction path, sometimes around an island of unexcitable or slowly conducting tissue. The wave of depolarization becomes self-sustaining, leading to a tachyarrhythmia.

One of the most serious arrhythmias is ventricular fibrillation (VF), which is associated with chaotic and rapid twitching of the ventricles without any effective pumping of blood through the heart. This may lead to cardiac arrest (i.e., failure of the heart to circulate blood around the body effectively) and death if not promptly treated. Ventricular fibrillation occurs because of a reentrant conduction pattern in the ventricles (as a parallel, reentry in the atria may lead to atrial flutter and atrial fibrillation).

Arrhythmias may be treated by drugs that raise or lower the heart rate, or by other more invasive interventions such as ablation of reentry pathway tissue. A persistent bradyarrhythmia may be treated by the implantation of an artificial pacemaker. An artificial pacemaker applies electrical stimulation to induce depolarization at a desired rate, preempting the heart's (too slow) natural pacemaker.

A tachyarrhythmia caused by reentry may be an emergency situation since it may lead to cardiac arrest. Such a condition is treated by the application of an electrical current across the chest. This depolarizes the entire myocardium so that reentrant patterns are interrupted. The heart's natural pacemaker then resumes control of heart rhythm. This technique is called *defibrillation* in the case of ventricular fibrillation. In other cases, the discharge has to be synchronized with the QRS complex in order to avoid *inducing* ventricular fibrillation; in these cases, this technique is called *cardioversion.*

Patients at high risk of tachyarrhythmias may receive an implantable cardioverter-defibrillator (ICD). This is a device implanted within the thoracic cavity, with leads to the ventricles, that may detect aberrant heart rhythms and apply electrical shocks to restore normal rhythm.

2.5 Post-ACS Risk Stratification

Since patients who experience ACS remain at an elevated risk of death, even after receiving treatment [27], post-ACS risk stratification is an important clinical step in determining which patients should be monitored and treated more (or less) aggressively. This section provides background information on post-ACS risk stratification methods. We consider the TIMI risk score (TRS), echocardiography, and long-term ECG-based techniques.

The TRS [106, 107, 108] provides a general assessment of risk based on clinical variables that can easily be obtained at the time of admission. The variables considered by the TRS represent seven significant independent predictors of risk. Echocardiography is a technique for imaging the heart using ultrasound; it yields information about blood flow in the heart as well as the shape of the heart. As described in Section 2.3, ECG data may also be used to diagnose the severity of an ACS at the time of presentation and is typically used to guide immediate treatment.

Each of the techniques considered here incorporates some information about a patient and yields a number that can be used to estimate the patient's risk. For example, higher values of the TRS are associated with higher risk. We evaluate the utility of these risk stratification techniques using two metrics. The c-statistic, or area under the receiver operating characteristic (ROC) [116], identifies the degree to which progressively higher values of a continuous variable are associated with an increased risk of adverse events. The Cox proportional hazards regression model, in contrast, estimates the relative ratio of the instantaneous rate of death (i.e., the hazard ratio) between different groups defined by a discrete variable. This may be useful when treatments are chosen based on a dichotomized value of a particular variable, e.g., if its value is above or less than some threshold value, or if the risk variable is not continuous and can only take a small number of discrete values (such as the TRS).

2.5.1 TIMI Risk Score

The TIMI risk score (TRS) [106, 107, 108] is a simple risk stratification technique that incorporates clinical variables easily acquired at the time of admission. It can therefore be used in triage and immediate decision-making with regard to treatment options. The GRACE [117] and PURSUIT [118] risk scores perform similar functions.

The TRS considers the following binary predictor variables:

- Age 65 years or older
- At least 3 risk factors for coronary artery disease among the following: hypertension, hypercholesterolemia, diabetes, family history of coronary artery disease, or being a current smoker
- Prior diagnosed coronary stenosis (narrowing of an artery) of 50% or more
- ST-segment deviation on ECG at presentation
- Severe anginal symptoms (at least 2 anginal events in prior 24 hours)
- Use of aspirin in prior 7 days
- Elevated serum cardiac markers (CK-MB or troponin)

One point is accrued for each variable that is observed, and the TIMI risk score is the total number of points (between 0 and 7). The set of variables was obtained by selecting independent prognostic variables from a set of 12 prospective clinical variables after a multivariate logistic regression [106]. The other 5 variables that were considered but not included in the TRS were: prior MI, prior coronary artery bypass graft (CABG), prior angioplasty (PTCA), prior history of congestive heart failure, and use of IV heparin within 24 hours of enrollment.

The TIMI 11B and ESSENCE trials [106] showed that a higher TRS is associated with higher rates of adverse events, defined as death, MI, or severe recurrent ischemia, in the 14 days following the initial event. The TIMI risk score has also been shown to be useful in risk stratification of patients over longer follow up periods [108, 109].

2.5.2 Echocardiography

Echocardiography (often referred to as simply "echo") is the use of ultrasound techniques to create an image of the heart. An echocardiogram can yield structural information about the heart and its valves, and information about blood flow through the heart. Magnetic resonance imaging (MRI), nuclear imaging, and angiography can provide some of the same information.

An echocardiogram is frequently used to assess left ventricular function [119]. The left ventricle is the largest chamber of the heart and is responsible for pumping oxygenated blood to the body. This leads to left ventricular function being critically important to the health of the body. If the myocardium is damaged, for example post-MI, the left ventricle may be unable to pump out sufficient blood. This may lead to symptoms of congestive heart failure, and has been shown to be strongly associated with fatal arrhythmias [120].

One measure of left ventricular function is the left ventricular ejection fraction (LVEF), i.e., the fraction of the blood volume ejected from the left ventricle during systole (the contraction phase of the heartbeat) relative to the volume present at the end of diastole (the relaxation phase of the heartbeat). An echocardiogram may be used to estimate the LVEF as:

$$\text{LVEF} \equiv \frac{(\text{LV volume before systole}) - (\text{LV volume after systole})}{(\text{LV volume before systole})} \tag{2.1}$$

A healthy heart has an LVEF of between 0.55 and 0.75 [95]. Patients with an LVEF of below 0.40 are considered as having significant left ventricular dysfunction [97]. The results of the MADIT II trial [120] strongly suggest that patients with LVEF less than 0.30 should have defibrillators implanted despite the risk of these invasive procedures.

2.5.3 Long Term ECG Techniques

A variety of methods have been proposed that assess risk based on automated analysis of long-term ECG data collected in the hours or days following admission. Such data is routinely collected during a patient's stay and therefore these additional risk assessments can be obtained at almost no additional cost. We discuss three ECG-based methods that have been proposed in the literature: heart rate variability (HRV) [93, 111], heart rate turbulence (HRT) [104], and deceleration capacity (DC) [90]. Each of these measures has been shown to correlate with risk of various adverse events in the period following an ACS. In subsequent parts of the thesis, we compare our methods to use information in long-term ECG signals with these existing ECG-based risk variables.

One additional long-term ECG-based risk stratification technique, T-wave alternans (TWA) [115], has also received some attention. However, evaluating TWA requires the use of specialized equipment and requires patients to complete specific maneuvers in order to elevate their heart rate. Unlike the other long-term ECG risk measures we consider, TWA cannot be computed using regular Holter monitor data. It is unlikely that TWA (in its current form) could be used widely for risk stratification in general populations, and as such we do not consider it further in this thesis.

Heart Rate Variability

The class of ECG-based risk stratification techniques that has been discussed most extensively in the literature is based on measurements of heart rate variability (HRV) [93, 111]. The theory underlying HRV-based techniques is that in healthy people, the body should continuously compensate for changes in oxygen demand, by changing the heart rate. The heart rate should also changes as a result of physiological phenomena such as respiratory sinus arrhythmia [95]. A heart rate that changes little suggests that the heart or its control systems are not actively responding to stimuli. HRV-based measures attempt to quantify the change in a patient's instantaneous heart rate over a period of monitoring in order to yield an estimate of risk.

Heart rate is primarily modulated by the autonomic nervous system, which comprises the the sympathetic and parasympathetic nervous systems. The parasympathetic nervous system's effects on heart rate are mediated by the release of acetylcholine by the vagus nerve, which lowers the heart rate. The sympathetic nervous system's effects are mediated by the release of epinephrine and norepinephrine, which raise heart rate. Decreased vagal or parasympathetic modulation (i.e. reduced down-regulation of heart rate) is thought to be strongly linked to increased risk of death [98, 99]. One possible explanation for this is that reduced down-regulation corresponds to an increase in heart rate, which although useful in maintaining a steady blood supply, further imposes stress on heart muscle already affected by ischemia or infarction. However, there is little consensus on whether low HRV is simply a correlate of poor outcomes, or whether it is part of some mechanism that leads to arrhythmias [93].

In general, HRV-based techniques first compute the sequence of intervals between heartbeats, which may be determined from ECG tracings. These are typically obtained by counting from one QRS complex to the next [93] since the QRS complex is the most prominent feature of a heartbeat. Abnormal beats are ignored, since the focus of HRV is to study how the nervous system modulates heart rate. While abnormal beats change the heart rate, these changes are the result of a different physiological phenomenon (e.g., the presence of abnormal beat foci) and are ignored so as not to be confused with heart rate changes due to impulses from the nervous system. Since only heartbeats resulting from normal depolarization of the SA node are considered, the sequence of R-wave to R-wave intervals studied for HRV analysis is termed the NN (for normal-to-normal) series. One of a number of methods is then used to summarize this series with a single number indicating the amount of heart rate variability. These HRV measures can be roughly divided into time domain, frequency domain, and nonlinear measures. [93] provides a more complete overview of HRV metrics.

Time domain HRV methods give a measure of total variation in heart rate. Commonly considered time domain HRV metrics include SDNN (standard deviation of NN intervals) and SDANN (standard deviation of mean NN interval over five-minute windows of the recording). Other time domain measures include:

- ASDNN, the mean of the standard deviation of NN intervals within five-minute windows.
- RMSSD, the root-mean-square of differences of successive NN intervals.
- HRVI (HRV triangular index), the maximum number of items in a single bin in a histogram of NN intervals (using a standard bin width of 1/128 s), divided by the total number of NN intervals.
- pNN50, the fraction of differences of successive NN intervals that exceeded 50 ms.

Frequency domain HRV methods rely on the fact that vagal and sympathetic activity are mediated by biochemical pathways associated with different time scales [93]. In particular, acetylcholine (which mediates vagal activity) is faster acting than epinephrine and norepinephrine (which mediate sympathetic activity). As a result, it is believed that changes in heart rate in the high frequency (HF) range (0.15-0.40 Hz) correspond to vagal activity, while changes in heart rate in the low frequency (LF) range (0.04-0.15 Hz) correspond to sympathetic activity. There is, however, considerable disagreement as to the specific phenomena measured by these bands [5, 6].

One of the most commonly used frequency domain metrics, LF/HF, is defined as the ratio of the total power at LF and HF frequencies in the power spectral density (PSD) of the NN series. The PSD of the NN series is usually measured using the Lomb-Scargle periodogram [89], which is designed to estimate the frequency content of a signal that is sampled at irregular intervals. This makes it well suited for the NN series, where samples are often irregularly spaced due to the removal of noisy parts of the ECG signal and abnormal beats. The LF/HF ratio is computed for 5-minute windows, as in [93], and the median value across windows is used as the LF/HF value for that patient. Patients with low HRV-LF/HF are considered to be at risk.

In our experiments, we found that HRV-LF/HF performed better at identifying patients at high risk of death post-ACS than any of the time domain metrics. These results are consistent with earlier findings reported by the Framingham Heart Study [101]. Frequency-based methods may be more robust, in general, because they focus on specific physiologically relevant frequencies in the NN series, and ignore artifacts at other frequencies.

Heart Rate Turbulence

Heart rate turbulence (HRT) [90] is related to HRV in that it studies the autonomic tone of patients. HRT studies the return to equilibrium of the heart rate after a premature ventricular contraction (PVC). Typically, following a PVC there is a brief speed-up in heart rate following by a slow decrease back to the baseline rate. This corresponds to the "turbulence" in the heart rate and is present in patients with a healthy autonomic nervous system.

HRT is essentially a baroreflex phenomenon. When a PVC interrupts the normal cardiac cycle, the ventricles have not had time to fill to their normal level, resulting in a weaker pulse. This triggers the homeostatic mechanisms that compensate by increasing heart rate. This compensatory increase in heart rate causes blood pressure to overcompensate and active the baroreflex in reverse.

HRT is quantified using two metrics: turbulence onset (TO) and turbulence slope (TS). The turbulence onset is a measurement of the acceleration in HR and is calculated based on the two RR intervals preceding the PVC and the two RR intervals immediately following the PVC.

$$TO = \frac{(RR_1 + RR_2) - (RR_{-1} + RR_{-2})}{(RR_{-1} + RR_{-2})} x100\% \tag{2.2}$$

TS is measured as the steepest slope of the linear regression line for each sequence of five consecutive RR intervals. High risk patients have either $TO \geq 0$ or $TS \leq 2.5$.

Patients with both these findings are at highest risk.

HRT was evaluated prospectively in a study on 1137 post acute MI patients [90]. Patients were categorized as either 0, 1 or 2 based on the outcome of the HRT test (0 corresponding to no risk findings and 2 corresponding to both risk findings) and followed for an average of 22 months. During this period 70 all-cause deaths occurred. On multivariate analysis in this patient population, HRT category 2 (i.e., having both $TO \geq 0$ and $TS \leq 2.5$) was found to be a stronger predictor of death than age, a history of diabetes, and LVEF.

Deceleration Capacity

Deceleration capacity (DC) is an extension of work on heart rate turbulence [104]. Like HRV, DC attempts to measure impaired vagal modulation of heart rate, which is believed to be associated with high risk. The theory underlying the DC technique is that vagal activity can be distinguished from sympathetic activity because vagal activation causes heart rate deceleration while the sympathetic nervous system causes heart rate acceleration [90].

To compute DC, we begin with the RR interval sequence $RR[n]$ and search for *anchors*, i.e., RR intervals that are longer than the ones preceding them. Denoting the index of the ith anchor as n_i and the total number of anchors by N, we define $X[n]$ as the average RR interval length around each anchor. This is measured by:

$$X[n] = \frac{1}{N}\sum_{i=1}^{N} RR[n_i + n] \tag{2.3}$$

DC is then computed from this information as:

$$\mathrm{DC} \equiv \frac{(X[0] + X[1]) - (X[-1] + X[-2])}{4}. \tag{2.4}$$

Roughly speaking, DC measures the magnitude of the typical beat-to-beat deceleration. The hypothesis underlying this work is that impaired deceleration corresponds to a heart that is unresponsive to vagal stimulation.

In 2006, a cohort study investigating DC as a predictor of mortality after MI showed that $DC > 4.5$ was indicative of an extremely low risk of death, while $DC \leq 2.5$ predicted high risk for mortality even in patients with preserved LVEF [90]. The cohort study developed DC cut-off values using data from 1455 patients from a post-infarction study in Munich. These cutoff values were then prospectively tested on a total of 1256 patients from both London, UK (656) and Oulu, Finland(600).

2.6 Summary

In this chapter, we reviewed clinical background on the heart, electrocardiogram, acute coronary syndromes, arrhythmias, and existing risk stratification methods. With this context in place, we now present our work on developing novel computational methods to identify high risk patients with cardiovascular disease. We focus, in particular, on risk stratification post-ACS and provide data in subsequent chapters showing how our work can improve upon the existing practice of medicine through traditional risk assessment techniques.

Chapter 3

Morphologic Variability

In this chapter, we present our work on morphologic variability (MV). MV is a measure of the amount of subtle variability in the shape of signals over long periods. The intuition underlying this work is that measuring variability in physiological signals provides useful prognostic information about the generative system. The presence of too little variability may suggest that the generative system is unresponsive and fails to react to external stresses. Conversely, the presence of too much variability may indicate a generative system that is unstable and is unable to settle in a happy medium of repeatable function.

In the context of cardiac disease, our hypothesis is that increased MV is associated with increased instability in the underlying physiological system, i.e., the heart or myocardium. Measuring subtle variations in the shape of the ECG signal is not straightforward. Key technical problems include detecting small changes in the presence of relatively large and variable time-skew, and finding techniques to summarize these changes across different time scales.

We show how these challenges can be addressed to produce a robust and powerful risk stratification tool. The results of our studies on MV, which we will discuss shortly, suggest that high MV is strongly associated with cardiovascular mortality and sudden cardiac death. This holds true even after adjusting for existing risk stratification approaches. In fact, there is a strong association between high MV and death even in patients who are missed by widely used risk assessment techniques such as echocardiography. This data indicates that MV is not only independent of other clinical measures, but moreover, it may potentially address cases where these other methods fail.

In what follows, Section 3.1 describes the pathophysiological basis for our work on morphologic variability and presents a hypothesis for how unstable bifurcations in the myocardium can lead to variability in the shape of ECG signals. Section 3.2 then details the process through which MV is measured, including a dynamic time-warping (DTW) algorithm for comparing the morphology of entire heartbeats. We present an evaluation of MV on ECG signals from post-ACS patients in Section 3.3. Finally, we show how MV can be used in settings beyond risk stratification for ACS in Section 3.4.

3.1 Pathophysiology of Morphologic Variability

In a stationary and homogeneous myocardial conducting system, the activated pathways through excitable cells are usually similar for consecutive beats. However, in the presence of ischemia, the conducting system has multiple irregular islands of severely depressed and unexcitable myocardium [78] that leads to discontinuous electrophysiological characteristics [81]. The presence of several possible adjacent pathways that can invade the non-functioning area leads to variations in the spatial direction of the invading vector [75]. Measured electrical activity in this phase can best be described in probabilistic terms because of beat-to-beat activation and repolarization variability, stemming from subtle unstable conduction bifurcations. Furthermore, propagation of a beat may be dependent on the route of propagation of the previous beat. The overall effect of such minor conduction inhomogeneities is not well understood, but it is possible that they correlate with myocardial electrical instability and have potentially predictive value for ventricular arrhythmias [75] or other adverse events.

Our pathophysiological hypothesis for MV is illustrated in Figures 3-1 to 3-4. A healthy myocardium (Figure 3-1) conducts electrical impulses smoothly. However, if parts of the myocardium are infarcted or ischemic (Figure 3-2), then the electrical

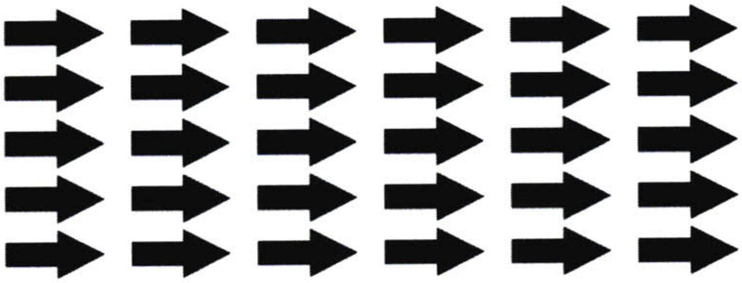


Figure 3-1: Healthy myocardium.

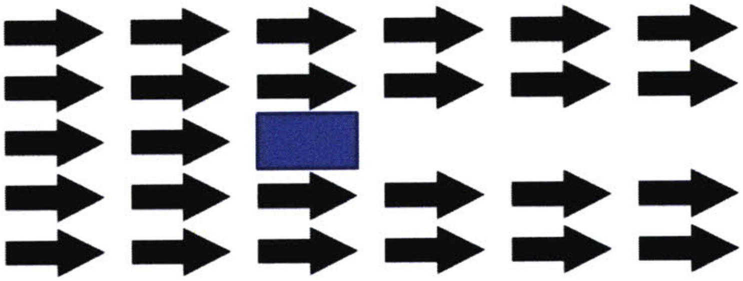


Figure 3-2: Myocardium with islands of non-conducting or slowly conducting tissue.

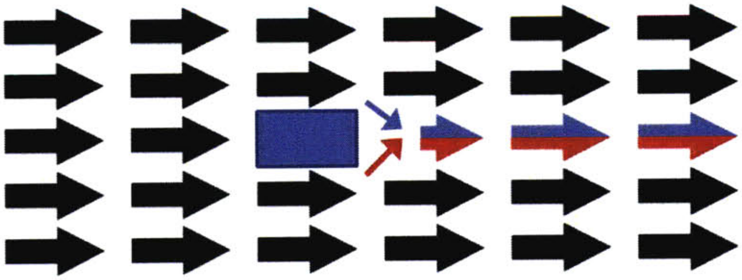


Figure 3-3: Race conditions leading to variability in ECG morphology.

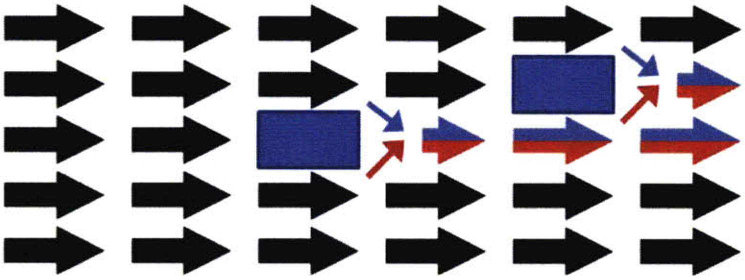


Figure 3-4: Time-varying islands of non-conducting or slowly conducting tissue.

impulse cannot propagate through this tissue and must pass around it. This leads to a race condition (Figure 3-3), where multiple advancing wavefronts from adjacent healthy tissue compete to polarize myocardium beyond the diseased region. The outcome of the race condition is probabilistic, and as the specific wavefront that first propagates around the infarcted or ischemic tissue changes from beat to beat, the path of the electrical impulse through the heart and the ECG morphology measuring the electrical field of the heart also changes from beat to beat. If the disease process itself is time-varying (Figure 3-4), e.g., due to intermittent coronary artery spasm, then this may cause the number of non-conducting or slowly conducting islands of tissue to further change over time. This represents an added source of variability in the shape of the ECG signal.

This variability can be quite subtle in practice. As an example of this, consider the two tracings shown in Figure 3-5. While it is hard to visually determine if one of the patients has more beat to beat variability in morphology, a significant difference is found computationally. The ECG tracing for the patient on the left has four times the morphologic variability of the patient on the right.

3.2 Measuring Morphologic Variability

The overall system for measuring MV is shown in Figure 3-6.

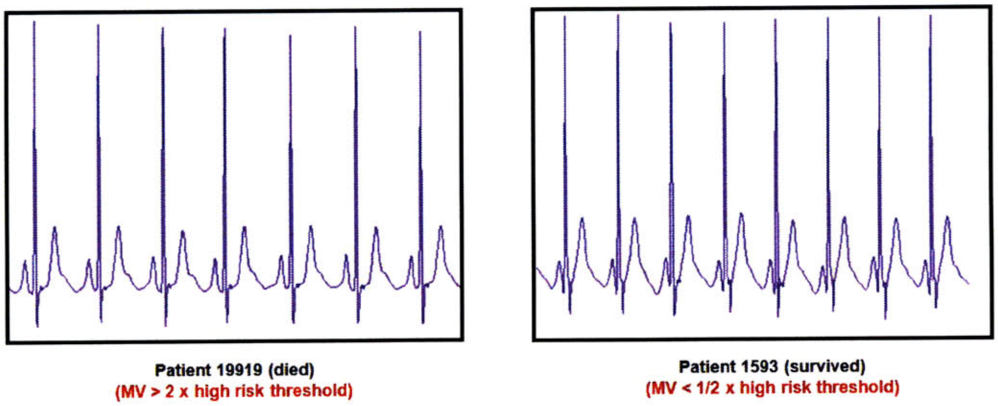

Patient 19919 (died)
(MV > 2 x high risk threshold)
Patient 1593 (survived)
(MV < 1/2 x high risk threshold)

Figure 3-5: ECG tracings from two patients.

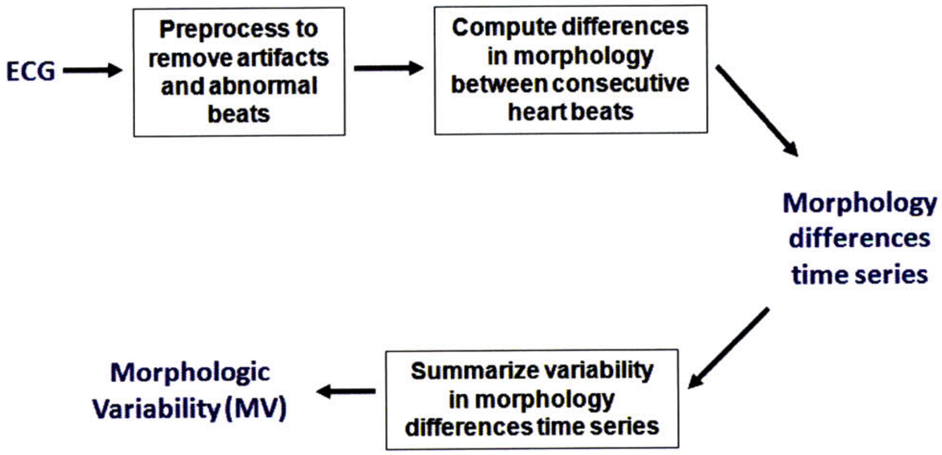

ECG
Preprocess to remove artifacts and abnormal beats
Compute differences in morphology between consecutive heart beats
Morphology differences time series
Summarize variability in morphology differences time series
Morphologic Variability (MV)

Figure 3-6: System for measuring morphologic variability.

3.2.1 ECG Signal Preprocessing

The process of analyzing ECG morphology is more sensitive to noise than techniques focusing exclusively on the heart rate. This is because the heart rate can often be estimated robustly, even in the presence of significant amounts of noise, by searching for high amplitude R-waves in the signal. In contrast, characterizing the morphology requires using information even from those parts of the ECG that are low amplitude and where small amounts of noise can significantly affect the signal-to-noise ratio. To minimize this effect, we employ techniques for noise removal and for automated signal rejection.

Noise removal is carried out in three steps over the length of the entire signal. Baseline wander is first removed by subtracting an estimate of the wander obtained by median filtering the original ECG signal as described in [76]. The ECG signal is then filtered using wavelet denoising with a soft-threshold [77]. Finally, sensitivity to calibration errors is decreased by normalizing the entire ECG signal by the mean R-wave amplitude.

While the noise removal steps help remove artifacts commonly encountered in long-term electrocardiographic records, the signal rejection process is designed to remove segments of the ECG signal where the signal-to-noise ratio is sufficiently low that meaningful analysis of the morphology is challenging even after noise removal. Such regions are typically dominated by artifacts unrelated to cardiac activity but that have similar spectral characteristics to the ECG signal, e.g., segments recorded during periods when there was substantial muscle artifact.

The process of signal rejection proceeds in two stages. Parts of the ECG signal with a low signal quality index [82] are first identified by combining four analysis methods: disagreement between multiple beat detection algorithms on a single ECG lead, disagreement between the same beat detection algorithm on different ECG leads, the kurtosis of a segment of ECG, and the ratio of power in the spectral distribution of a given ECG segment between 5-14 Hz and 5-50 Hz. In our work, we use the Physionet SQI package implementation [82] to carry out these analyses and automatically remove parts of the ECG signal with a low signal quality index. The remaining data is then divided into half hour windows, and the standard deviation of the R-waves during each half hour window is calculated. We discard any window with a standard deviation greater than 0.2887. Given the earlier normalization of the ECG signal, under a conservative model that allows the R-wave amplitude to uniformly vary between 0.5 and 1.5 every beat (i.e., up to 50% of its mean amplitude), we expect the standard deviation of the R-wave amplitudes to be less than 0.2887 for any half hour window. This heuristic identifies windows that are likely corrupted by significant non-physiological additive noise, and where the morphology of the ECG cannot be meaningfully analyzed.

3.2.2 ECG Segmentation and Removal of Ectopy

To segment the ECG signal into beats, we use two open-source QRS detection algorithms with different noise sensitivities. The first of these makes use of digital filtering and integration [79] and has been shown to achieve a sensitivity of 99.69%, while the second is based on a length transform after filtering [86] and has a sensitivity of 99.65%. Both techniques have a positive predictivity of 99.77%. QRS complexes were marked only at locations where these algorithms agreed.

Prior to further analysis, we also remove ectopic parts of the signals in a fully automated manner. This is done using the beat classification algorithm of [80] present in the Physionet SQI package. The beat classification algorithm characterizes each beat by a number of features such as width, amplitude and RR interval, and then compares it to previously detected beat types to assign it a label.

The decision to remove ectopic beats is motivated by the fact that MV is designed to measure variations in morphology arising from unstable bifurcations in the myocardium. While ectopic beats have changes in morphology that are suggestive of abnormalities, the source of these abnormalities corresponds to a different underlying phenomenon (i.e., abnormal impulse formation as opposed to abnormalities in impulse propagation). For measuring MV, we therefore remove ectopic beats from analysis.

Due to the removal of ectopic beats, MV is measured on mainly "normal looking" ECG. We believe that analyzing this data is one of the strengths of our work, i.e., MV focuses on precisely the same data that is often considered to be clinically uninteresting yet contains much valuable information. As the results of our evaluations in Section 3.3 show, this seemingly unintuitive decision to study variation in "normal looking" ECG morphology allows us to discover important new information that is independent of other widely used risk metrics.

3.2.3 Comparing ECG Beats

We develop a metric that quantifies how the ECG morphology of two beats differs. The simplest way to calculate this energy difference is to subtract the samples of one beat from another. However, if samples are compared based strictly on their distance from the start of the beat, this process may end up computing the differences between samples associated with different waves or intervals. For example, consider the two heart beats depicted in Figure 3-7. In the drawing on the left, samples are aligned based on their distance from the onset of the P-wave. Vertical lines connect corresponding samples from the beat colored red to the beat colored blue. If samples are compared strictly on their distance from the start of the beat, this process may end up computing the difference between unrelated parts of the two beats. In the drawing on the left, samples are aligned based on their distance from the onset of the P-wave. One consequence of this is that samples that are part of the T-wave of the top beat are compared with samples not associated with the T-wave of the bottom beat. A measure computed this way will not reflect differences in the shapes of the T-waves of adjacent beats.

We use a variant of dynamic time-warping (DTW) [85] to align samples that correspond to the same underlying physiological activity. As depicted in the drawing on the right side of Figure 3-7, this can lead to aligning a single sample in one beat with multiple samples in another beat. The algorithm uses dynamic programming to search for an alignment that minimizes the overall distortion. Distortion is measured using the method described in [87], which captures differences in both amplitude and

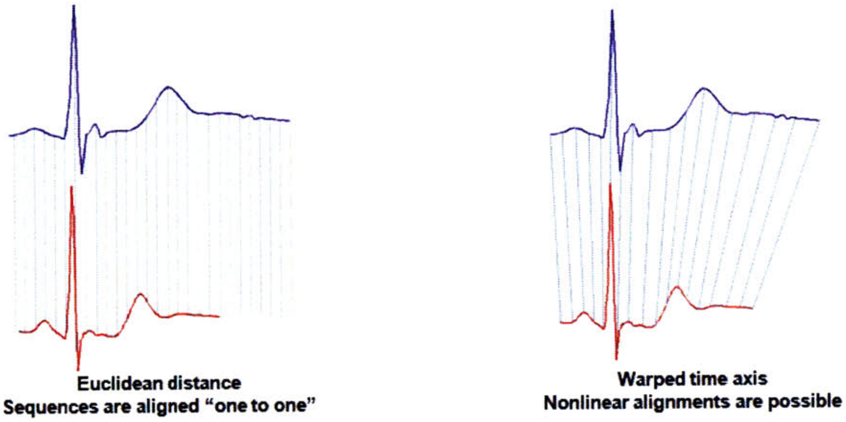

Euclidean distance
Sequences are aligned "one to one"
Warped time axis
Nonlinear alignments are possible

Figure 3-7: Alignment of beats by dynamic time-warping.

timing of ECG waves.

More precisely, given two beats, x_1 and x_2, of length l_1 and l_2 respectively, DTW produces the optimal alignment of the two sequences by first constructing an l_1-by-l_2 distance matrix d. Each entry (i,j) in this matrix d represents the square of the difference between samples $x_1[i]$ and $x_2[j]$. A particular alignment then corresponds to a path, f through the distance matrix of the form:

$$\phi(k) = (\phi_1(k), \phi_2(k)),\ 1 \leq k \leq K \tag{3.1}$$

where f_1 and f_2 represent row and column indices into the distance matrix, and K is the alignment length. Any feasible path must obey the endpoint constraints:

$$\phi_1(1) = \phi_2(1) = 1 \tag{3.2}$$

$$\phi_1(K) = l_1 \tag{3.3}$$

$$\phi_2(K) = l_2 \tag{3.4}$$

as well as the continuity and monotonicity conditions:

$$\phi_1(k+1) \leq \phi_1(k) + 1 \tag{3.5}$$

$$\phi_2(k+1) \leq \phi_2(k) + 1 \tag{3.6}$$

$$\phi_1(k+1) \geq \phi_1(k) \tag{3.7}$$

$$\phi_2(k+1) \leq \phi_2(k) \tag{3.8}$$

The optimal alignment produced by DTW minimizes the overall cost:

$$C(x_1, x_2) = \min_{\phi} \; C_\phi(x_1, x_2) \tag{3.9}$$

where C_ϕ is the total cost of the alignment path f and is defined as:

$$C_\phi(x_1, x_2) = \sum_{k=1}^{K} d(x_1[\phi_1(k)], x_2[\phi_2(k)]) \tag{3.10}$$

The search for the optimal path is carried out in an efficient manner using a dynamic programming algorithm derived from the following recurrence for the cumulative path distance, $\gamma(i, j)$, and the distance matrix d:

$$\gamma(i,j) = d(i,j) + \min \left\{ \begin{array}{c} \gamma(i-1, j-1) \\ \gamma(i-1, j) \\ \gamma(i, j-1) \end{array} \right\} \tag{3.11}$$

The final energy difference between the two beats x_1 and x_2, is given by the cost of their optimal alignment, which depends on the amplitude differences between the two signals and the length, K, of the alignment (which increases if the two signals differ in their timing characteristics). In a typical formulation of DTW, this difference is divided by K to remove the dependence of the cost on the length of the original observations. A problem with applying this correction in our context is that some paths are long not because the segments to be aligned are long, but rather because they differ in length. In these cases, dividing by K is inappropriate since a difference in the length of a beats (or of parts of beats) often provides diagnostic information that is complementary to the information provided by the morphology. Consequently, in our algorithm we omit the division by K.

A further modification to traditional DTW in our work is that 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 d, 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 in [83]. It 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 (which is also shown graphically in Figure 3-8):

$$\gamma(i,j) = d(i,j) + \min \left\{ \begin{array}{c} \gamma(i-1,j-1) \\ d(i-1,j) + \gamma(i-2,j-1) \\ d(i-1,j) + d(i-2,j) + \gamma(i-3,j-1) \\ d(i,j-1) + \gamma(i-1,j-2) \\ d(i,j-1) + d(i,j-2) + \gamma(i-1,j-3) \end{array} \right\} \tag{3.12}$$

3.2.4 Morphologic Distance (MD) Time Series

Using the process described above, we can transform the original ECG signal from a sequence of beats to a sequence of energy differences between consecutive pairs of beats. We call the resulting time series the morphologic distance (MD) time series for the patient. This new signal, comprising pair-wise, time-aligned energy differences between beats, is then smoothed using a median filter of length 8. The median filtering process addresses noisy and ectopic heart beats that may have passed through the

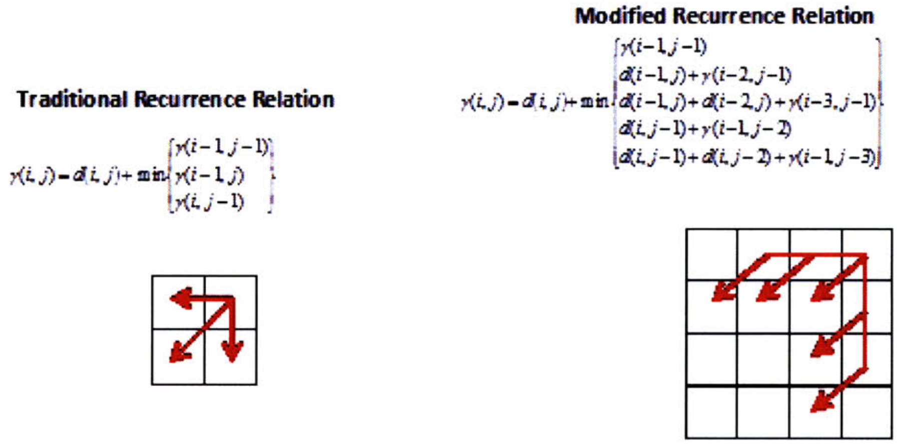

Traditional Recurrence Relation
$\gamma(i,j) = d(i,j) + \min\begin{Bmatrix} \gamma(i-1,j-1) \\ \gamma(i-1,j) \\ \gamma(i,j-1) \end{Bmatrix}$
Modified Recurrence Relation
$\gamma(i,j) = d(i,j) + \min\begin{Bmatrix} \gamma(i-1,j-1) \\ d(i-1,j)+\gamma(i-2,j-1) \\ d(i-1,j)+d(i-2,j)+\gamma(i-3,j-1) \\ d(i,j-1)+\gamma(i-1,j-2) \\ d(i,j-1)+d(i,j-2)+\gamma(i-1,j-3) \end{Bmatrix}$

Figure 3-8: Traditional and modified recurrence relation of dynamic time-warping.

earlier preprocessing stage and lead to high morphologic distances. The smoothing process is geared towards ensuring that high values in the MD time series correspond to locally persistent morphology changes, i.e., sustained differences in beat-to-beat morphology.

3.2.5 Spectral Energy of Morphologic Differences

We choose to summarize information in the MD time series in the frequency domain, rather than the time domain. This decision is based on a desire to reduce the influence of noise on MV measurements. In contrast to summing up information in the MD signal in the time domain, i.e., over all frequencies, our objective is to measure energy with a specific frequency range of the MD power spectral density. This frequency range is chosen to best distinguish between pathological variability, and variability due to noise. In this way, our approach is analogous to work on HRV (Section 2.5.3), where frequency domain measures provide an improvement over risk assessment using time domain measures.

We estimate the power spectral density of the MD time series using the Lomb-Scargle periodogram [89], which is well-suited to measure the spectral content of an irregularly sampled signal (such as the MD series which is unevenly sampled following the removal of noise and ectopy). The Lomb-Scargle periodogram takes into account both the signal value and the time of each sample. For a time series where the value $m[n]$ is sampled at time $t[n]$, the energy at frequency ω is estimated as:

$$P(\omega) = \frac{1}{2\sigma^2}\{\frac{\sum_n[(m[n]-\mu)\cos\omega(t[n]-\tau)]^2}{\sum_n \cos^2\omega(t[n]-\tau)} + \frac{\sum_n[(m[n]-\mu)\sin\omega(t[n]-\tau)]^2}{\sum_n \sin^2\omega(t[n]-\tau)} \tag{3.13}$$

where and s are the mean and variance of the $m[n]$, and t is defined as:

$$tan(2\omega\tau) = \frac{\sum_n \sin(2\omega t[n]}{\sum_n \cos(2\omega t[n])} \tag{3.14}$$

We note that when using the Lomb-Scargle periodogram both the value and time of the MD samples are supplied as inputs to the estimation process. Therefore, rather than measuring the power spectrum of a signal indexed by sample number, we measure the power spectrum of a signal (in this case the MD time series) that is indexed by absolute time. This allows us to estimate the power spectrum of beat to beat changes in morphology in terms of absolute frequencies (i.e., in terms of seconds rather than samples). This is useful because it allows us to directly relate the frequency at which changes occur in the MD signal to the pharmacokinetics of autonomic modulators such as acetylcholine, epinephrine and norepinephrine.

To distinguish between pathological variations in morphology and those resulting from noise or non-pathological physical effects, we investigated an absolute frequency range within the power spectral density of the MD time series that had maximal prognostic information. This was done by evaluating all possible frequency bands within 0.10.6 Hz with a granularity of 0.01 Hz in a training set of data. The range of $0.1 - 0.6$ Hz was based on the theory of how the vagal and sympathetic nervous systems modulate the heart through different biochemical pathways (Section 2.5.3). Within the rough range predicted by theory, we empirically derived the frequency range containing maximum diagnostic information.

We used ECG recordings from 764 patients in the DISPERSE2 TIMI33 study [88]

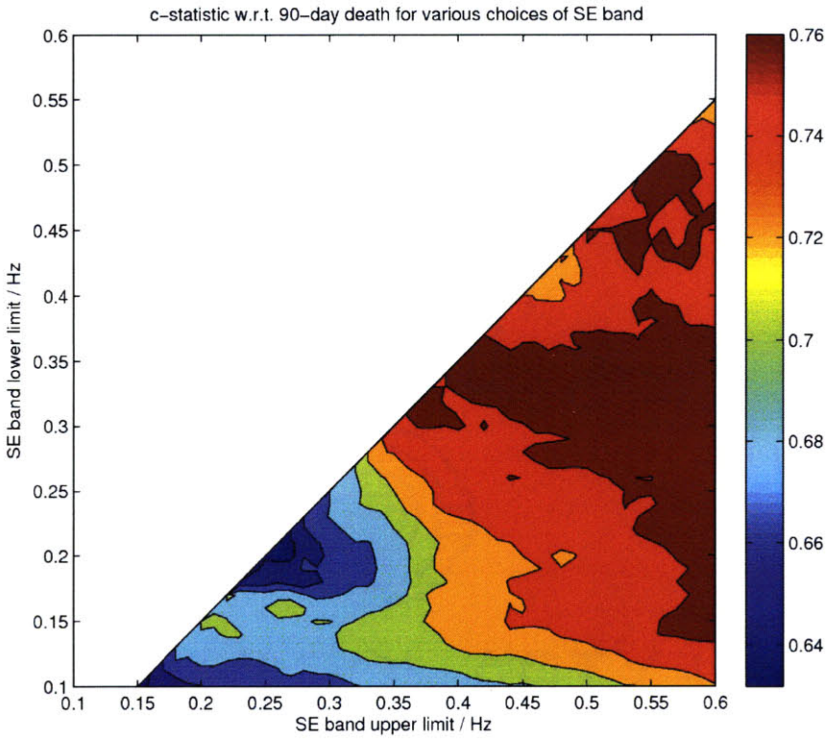


Figure 3-9: Heatmap showing c-statistic as a function of low- and high-frequency cutoffs. A maximum c-statistic value of 0.77 is obtained for the frequency band 0.300.55 Hz.

for this purpose. Patients in the DISPERSE2 study had 24 hour Holter ECG data recorded at 128 Hz within 48 hours of admission due to NSTEACS. 15 deaths were observed in this population over a follow-up period of 90 days. For each possible choice of prognostic frequencies, we computed energy in the MD power spectrum for all patients in the study population. These ranges were evaluated for their ability to discriminate between those patients who died and those who did not by measuring the corresponding c-statistic (i.e., the area under the receiver operating characteristic curve) [84]. As shown in Figure 3-9, the frequency range of 0.30-0.55 Hz yielded the maximum prognostic information in the study population and led to a training set c-statistic of 0.77. Based on this experiment, we define MV as the energy in the MD time series power spectral density between 0.300.55 Hz.

We measure MV over five minute windows, and aggregate information across windows for longer recordings. Specifically, given MV values for different five minute windows of ECG, the final MV for the patient is set to the 90th percentile of the five minute values. The choice of five minute windows and the 90th percentile are based on empirical evidence on patients in the DISPERSE2 TIMI33 study (a detailed discussion of these experiments can be found in [68]).

Consistent with existing risk stratification approaches, which dichotomize continuous variables into different risk groups with potentially distinct therapeutic consequences, we set the high risk cutoff for MV at 52.5 (i.e., values above 52.5 indicate patients at increased risk). This value was chosen as the highest quintile of MV in the DISPERSE2 TIMI33 population, in keeping with tradition [93].

3.3 Evaluation of Morphologic Variability

We developed and fine-tuned MV on training data from patients in the DISPERSE2 TIMI33 study. We then evaluated MV without any changes in a blind study on patients from a different trial, i.e., the MERLIN TIMI36 trial [94].

3.3.1 Methodology

The MERLIN TIMI36 trail was designed to compare the efficacy of ranolazine, a new anti-anginal drug, to matching placebo. The study enrolled 6560 patients within 48 hours of admission for NSTEACS. Full inclusion and exclusion criteria, as well as study procedures, have been previously published [94]. Patients in the MERLIN TIMI36 trial received standard medical and interventional therapy according to local practice guidelines, and were followed for up to 24 months (median follow-up of 348 days). Holter ECG recordings (Lifecard CF, DelMar Reynolds/Spacelabs, Issaqua, WA) were performed at 128 Hz in 6455 (98%) patients for a median duration of 6.8 days after randomization. Since all patients had Holter ECG for at least a day, we focused on the first 24 hours of ECG for evaluation.

MV was measured for each patient as described earlier. To assess the extent to which MV provided information beyond existing risk stratification techniques, we also included HRT, DC, LVEF and TRS in our study (i.e., all risk variables from Section 2.5 with the exception of HRV). The decision to leave out HRV was based on the fact that HRT and DC have been shown in recent studies to be better measures of impaired sympathovagal modulation of the heart than HRV [104, 90]. The use of these metrics, therefore, made the inclusion of HRV unnecessary as it does not add any useful information beyond HRT and DC for risk stratification.

We measured HRT and DC for each patient using the libRASCH software provided for research use by the inventors of the method (Technische Universitat Mnchen, Munich, Germany) [104, 90]. HRT was categorized based on the turbulence onset (TO) and turbulence slope (TS) as follows: HRT0 (TO$<$ 0 and TS$>$ 2.5), HRT1 (either TO=0 or TS=2.5ms), and HRT2 (TO=0 and TS=2.5ms). These categories were based on earlier results suggesting that the risk of death increases in a graded manner from patients with HRT0 to those with HRT2 [104]. DC was categorized as follows: category 0 (> 4.5), category 1 ($2.5 - 4.5$), and category 2 (< 2.5). These categories were based on earlier results suggesting that the risk of death increases in a graded manner from patients with DC$>$ 4.5 to those with DC$<$ 2.5 [90].

For LVEF and the TIMI risk score, we used the values already available as part of the MERLIN TIMI36 dataset. We divided patients into three LVEF categories, LVEF$<$ 30%, 30%$\leq$LVEF$<$40%, and LVEF$\geq$40%. The decision to categorize patients in this manner was based on a lack of consensus across different studies as to the right cutoff for LVEF dichotomization. Values between 30% [102] and 40% [103] have been used in different studies. We therefore decided to study patients with LVEF between 30% and 40% as a separate group. For the TRS, we categorized patients as low risk (TRS=1,2), moderate risk (TRS=3,4) and high risk (TRS=5,6,7), consistent with earlier work [94].

All statistical analyses were performed using Stata version 9.2 (StataCorp LP, College Station, TX). Hazard ratios (HR) and 95% confidence intervals (CIs) were estimated by use of a Cox proportional-hazards regression model. Event rates are presented as Kaplan-Meier failure rates at 52 weeks.

To address the issue of non-negligible amounts of correlation between the ECG variables, we did not evaluate the ECG variables simultaneously in a multivariate model. One problem caused by the presence of correlation, while trying to understand the interplay of predictors in multivariate models, is that the values of the regression coefficients may be distorted. Often, they are quite low and may fail to achieve statistical significance. This effect takes place because standard regression procedures assess each variable while controlling for all other predictors. At the time the procedure evaluates a given variable, other correlated variables may account for its variance. Another problem is that the standard errors of the regression weights of correlated predictors can be inflated, thereby enlarging their confidence intervals. More details on these and other issues can be found in [92].

As a result of this, instead of comparing the ECG variables collectively in a standard multivariate model, we carried out a different analysis. We assessed each ECG variable separately, and studied its usefulness while accounting for TRS and LVEF.

3.3.2 Results

After accounting for missing files, 4557 (71%) of the recordings were available for further analysis. There were no significant differences in the clinical characteristics of patients with and without available ECG data (Table 3.1). In the patient population with ECG recordings, there were 195 cardiovascular deaths (including 81 sudden cardiac deaths) and 347 myocardial infarctions during the follow-up period.

Table 3.2 presents the correlation between the different risk variables in the patient population with available data. The ECG risk variables had low to moderate levels of correlation with each other, and a low correlation with both LVEF ($R \leq 0.21$) and the TIMI risk score ($R \leq 0.18$).

Results of univariate and multivariate analysis for cardiovascular death are presented in Tables 3.3 and 3.4. We did not notice a statistically significant association between any of the ECG variables and cardiovascular death in patients with LVEF<40% (n=266, events=39). On univariate analysis, we found that MV was strongly associated with cardiovascular death over follow-up in patients with LVEF≥40%. These results continued to hold even after adjusting for the TRS, which comprises a

Table 3.1: Baseline clinical characteristics for patients with and without available data.

	Patients with Data (n=4557)	Patients without Data (n=1898)
Age, median (IQR), y	63 (55-72)	65 (56-72)
Female sex	35	36
BMI, median (IQR)	29 (25-31)	28 (26-32)
Diabetes Mellitus	34	35
Hypertension	73	75
Hyperlipidemia	67	68
Current smoker	26	23
Prior MI	33	37
Prior angina	55	58
Index event		
Unstable angina	47	50
NSTEMI	53	50

Table 3.2: Correlation between different risk variables following dichotomization.

	HRT	DC	MV	TRS	LVEF
HRT	1.00	0.39	0.22	0.16	0.15
DC		1.00	0.43	0.18	0.21
MV			1.00	0.10	0.13
TRS				1.00	0.13
LVEF					1.00

Table 3.3: Association of risk variables with cardiovascular death in patients with LVEF≥40% (HR=hazard ratio, CI=confidence interval, P=P value).

Parameter	HR	95% CI	P
HRT			
1 vs. 0	1.82	1.09–3.04	0.021
2 vs. 0	3.38	1.83–6.27	<0.001
DC			
1 vs. 0	2.42	1.51–3.88	<0.001
2 vs. 0	3.29	1.79–6.07	<0.001
MV>52.5	3.21	2.07–4.95	<0.001

variety of information related to the clinical characteristics of the patients, biomarkers, and medications. Similar results were obtained for HRT and DC, although MV had a higher hazard ratio than either of these metrics after adjusting for the TRS.

For the endpoint of sudden cardiac death, we obtained results that paralleled those for cardiovascular death (Tables 3.5 and 3.6). In particular, MV was strongly associated with sudden cardiac death in patients with LVEF≥40%, and these results were consistent even after adjusting for the TRS. In this case, however, neither DC nor HRT were associated with sudden cardiac death during follow-up after adjusting for the TRS. Similar to our results for cardiovascular death, we did not notice a statistically significant association between any of the ECG variables and sudden cardiac death in patients with LVEF<40% (n=266, events=18).

Finally, in the case of myocardial infarction, we did not see a significant association between MV and the endpoint during follow-up (Tables 3.7 and 3.8). In general, all three ECG risk variables performed poorly for this event although HRT and DC appear to have some promise. We did not notice a statistically significant association between any of the ECG variables and myocardial infarction in patients with LVEF<40% (n=266, events=36).

Kaplan-Meier curves for MV and the endpoints of cardiovascular mortality, sudden cardiac death, and myocardial infarction and presented in Figures 3-10 to 3-12.

Our results on the MERLIN TIMI36 data are quite encouraging and suggest that MV could play an important role in identifying high risk patients for both cardiovascular death and sudden cardiac death. In particular, we note that MV is strongly

Table 3.4: Association of risk variables with cardiovascular death in patients with LVEF≥40% after adjusting for the TIMI risk score (the TIMI risk score comprises the following predictors: age 65 years or older, at least 3 risk factors for coronary artery disease, prior coronary stenosis of 50% or more, ST-segment deviation on electrocardiogram at presentation, at least 2 anginal events in prior 24 hours, use of aspirin in prior 7 days, and elevated serum cardiac markers) (HR=hazard ratio, CI=confidence interval, P=P value).

Parameter	HR	95% CI	P
HRT			
1 vs. 0	1.60	0.96–2.67	0.073
2 vs. 0	2.65	1.42–4.97	0.002
DC			
1 vs. 0	2.11	1.32–3.39	0.002
2 vs. 0	2.58	1.40–4.77	0.002
MV>52.5	2.93	1.90–4.54	<0.001

Table 3.5: Association of risk variables with sudden cardiac death in patients with LVEF≥40% (HR=hazard ratio, CI=confidence interval, P=P value).

Parameter	HR	95% CI	P
HRT			
1 vs. 0	2.03	1.01–4.05	0.046
2 vs. 0	2.18	0.80–5.94	0.129
DC			
1 vs. 0	1.73	0.87–3.41	0.115
2 vs. 0	2.02	0.77–5.30	0.156
MV>52.5	2.44	1.28–4.66	0.007

Table 3.6: Association of risk variables with sudden cardiac death in patients with LVEF≥40% after adjusting for the TIMI risk score (the TIMI risk score comprises the following predictors: age 65 years or older, at least 3 risk factors for coronary artery disease, prior coronary stenosis of 50% or more, ST-segment deviation on electrocardiogram at presentation, at least 2 anginal events in prior 24 hours, use of aspirin in prior 7 days, and elevated serum cardiac markers) (HR=hazard ratio, CI=confidence interval, P=P value).

Parameter	HR	95% CI	P
HRT			
1 vs. 0	1.80	0.90–3.61	0.097
2 vs. 0	1.90	0.68–5.26	0.218
DC			
1 vs. 0	1.53	0.77–3.03	0.224
2 vs. 0	1.66	0.63–4.38	0.309
MV>52.5	2.27	1.19–4.32	0.013

Table 3.7: Association of risk variables with myocardial infarction in patients with LVEF≥40% (HR=hazard ratio, CI=confidence interval, P=P value).

Parameter	HR	95% CI	P
HRT			
1 vs. 0	1.69	1.21–2.36	0.002
2 vs. 0	1.86	1.14–3.04	0.014
DC			
1 vs. 0	1.66	1.20–2.28	0.002
2 vs. 0	1.28	0.75–2.20	0.365
MV>52.5	1.12	0.77–1.62	0.553

Table 3.8: Association of risk variables with myocardial infarction in patients with LVEF≥40% after adjusting for the TIMI risk score (the TIMI risk score comprises the following predictors: age 65 years or older, at least 3 risk factors for coronary artery disease, prior coronary stenosis of 50% or more, ST-segment deviation on electrocardiogram at presentation, at least 2 anginal events in prior 24 hours, use of aspirin in prior 7 days, and elevated serum cardiac markers) (HR=hazard ratio, CI=confidence interval, P=P value).

Parameter	HR	95% CI	P
HRT			
1 vs. 0	1.54	1.10–2.15	0.012
2 vs. 0	1.53	0.93–2.52	0.097
DC			
1 vs. 0	1.50	1.09–2.07	0.013
2 vs. 0	1.11	0.64–1.91	0.712
MV>52.5	1.04	0.72–1.51	0.824

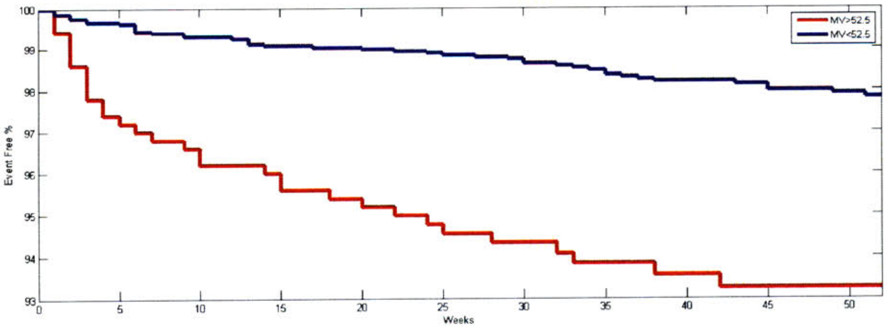


Figure 3-10: Kaplan-Meier survival curves for cardiovascular death in patients with LVEF≥40%.

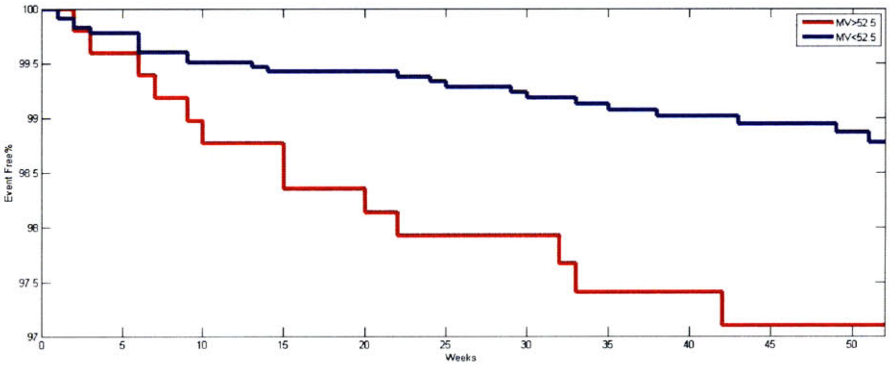


Figure 3-11: Kaplan-Meier survival curve for sudden cardiac death in patients with LVEF≥40%.

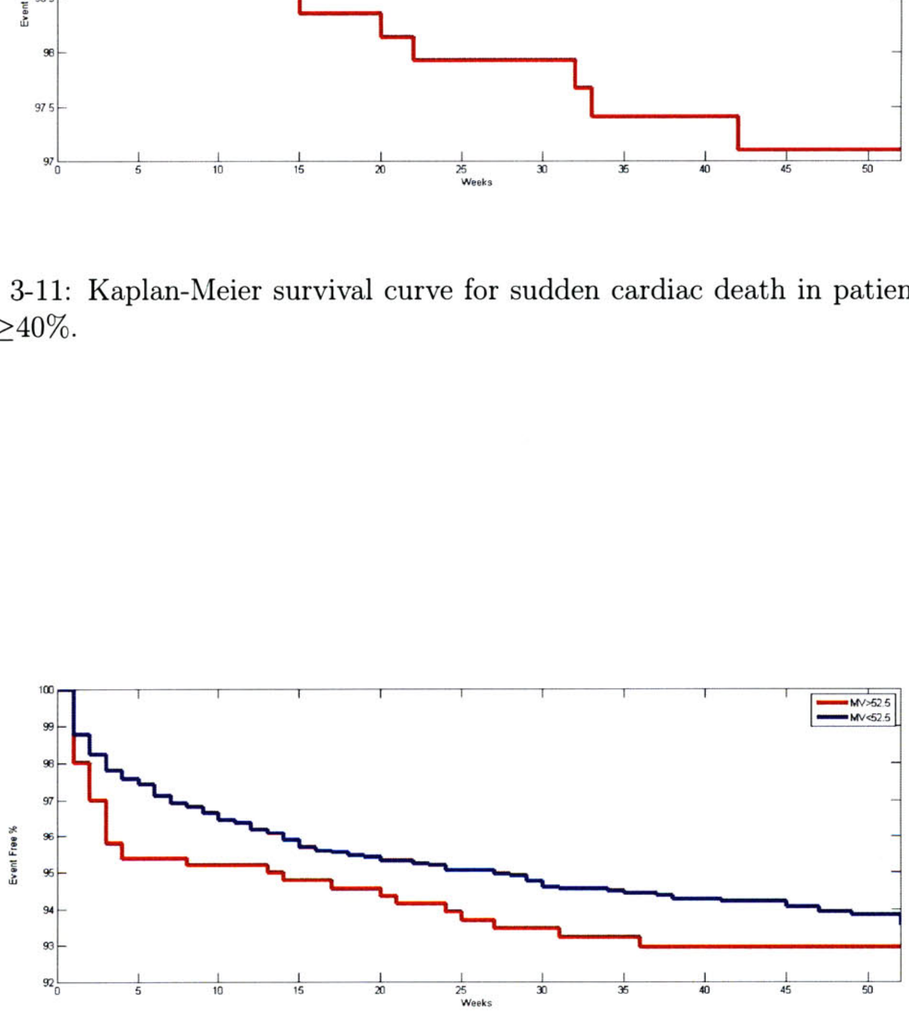


Figure 3-12: Kaplan-Meier survival curve for myocardial infarction in patients with LVEF≥40%.

associated with both these outcomes in patients with preserved LVEF. These results hold even after adjusting for the TIMI risk score.

The preserved LVEF patient group is particularly interesting for a variety of reasons. First, it represents the overwhelming majority of the patients (almost 90% of the patients in the MERLIN TIMI36 trial for whom LVEF was available had LVEF≥40%) and also the majority of the cardiovascular (68%) and sudden cardiac (69%) deaths that occurred during follow-up had preserved LVEF. Second, the preserved LVEF group is also the most challenging to risk stratify, since these patients appear to be low risk according to echocardiography. Third, the deaths that take place in patients who have low LVEF correspond to patients failing to respond to treatments once they have been determined to be high risk by echocardiography. In contrast, we speculate that the deaths that take place in patients who have preserved LVEF correspond to patients who were missed by echocardiography and might have befitted from more aggressive treatment. From this perspective, we believe that improved diagnosis may have a greater impact in patients with LVEF≥40% group.

Among patients already determined to be at high risk by echocardiography, we did not see an improvement with the use of any of the ECG risk variables. We believe that this result is in part due to statistical limitations, i.e., only a small minority of patients are categorized as high risk by echocardiography and the number of patients does not provide sufficient power for analysis. We also note that patients determined to be high risk by echocardiography may have received treatments that confound the evaluation process (e.g., some patients may have received ICDs).

Our results are also consistent for both cardiovascular death and sudden cardiac death. We consider this a particularly exciting aspect of our work. In contrast to death due to known cardiac causes, sudden cardiac death is poorly understood and widely considered as being harder to predict. Furthermore, treatment for sudden cardiac death generally corresponds to ICD implantation, which is both expensive and invasive. False positives therefore represent a huge inefficiency. At the same time, missing sudden cardiac death cases can also be disastrous. Therefore, errors in risk stratification for sudden cardiac death may be more costly than for death in general. We are therefore encouraged by the potential presented by MV to identify patients at high risk of sudden cardiac death.

Finally, we note that our evaluation is restricted to post-NSTEACS patients. This is a consequence of the availability of patient data. We believe that MV may be useful in other patient populations, including the diagnosis of patients without prior history of cardiac disease for *de novo* events. More testing is needed, however, to validate this hypothesis.

3.4 Other Applications of MV: Quantifying Treatment Effects

In addition to developing risk metrics that can be used to predict adverse outcomes, an associated goal of our work has been to explore the use of these metrics to quantify the impact of various treatments. We believe that this may help clinicians make better decisions about which therapies are working and which should be reconsidered.

In this context, we explored the use of MV and other ECG metrics to assess the impact of ranolazine (ranexa), a new cardiovascular drug that was recently proposed and formed the basis of the MERLIN trial [94]. Ranolazine failed to show any improvement in the survival of patients following NSTEACS during the trial, but is still widely used for symptomatic relief.

We studied differences in MV, HRT and DC between the ranolazine and placebo patients in MERLIN using the Wilcoxon rank sum test. Each ECG parameter was measured on the first 24 hours of data. We also compared the percentage of patients in each group considered to be high risk by the ECG metrics categorized as described earlier.

Treatment with ranolazine did not change HRT, but resulted in significantly lower MV and DC (Table 3.9), with a decrease in MV corresponding to *decreased* risk and a decrease in DC corresponding to *increased* risk. Similar results were seem in subsequent days. The percentage of patients at risk by ECG metrics in the ranolazine

Table 3.9: Characteristics of ECG variables in ranolazine and placebo group.

Parameter	Ranolazine Group (n=2255)	Placebo Group (n=2302)	P Value
DC	5.3 (3.9–6.9)	5.6 (4.0–7.6)	<0.001
MV	43.4 (29.6–46.4)	46.7 (30.3–50.0)	<0.001

group was 36.4% (DC) and 16.9% (MV) compared to the placebo group 31.0% (DC) and 21.7% (MV).

Our results suggest that ranolazine may potentially reduce myocardial electrical instability (as reflected by lower MV), while adversely dampening the autonomic responsiveness of the heart (resulting in lower DC).

More generally, the data from our experiment also motivates the use of computerized metrics to evaluate a broader set of therapies and to make personalized decisions on which treatments are appropriate for individual patients. Using metrics such as HRT, DC and MV, it might be possible to "troubleshoot" treatments and make precise, quantifiable statements about how they affect different physiological processes. This could play an important role in improving the existing practice of clinical trials by providing an inexpensive and quick means of supplementing the coarse survival data available through these trials.

3.5 Summary

In this chapter, we introduced the concept of morphologic variability. Our hypothesis was that subtle variability in the shape of signals over long periods, which is often confused for noise, contains valuable information about the underlying generative system. We motivate our work with a theory of how subtle changes in ECG morphology over long periods may be indicative of unstable bifurcations in the myocardium.

In addition to introducing the concept of morphologic variability, we also described a system to measure MV. We addressed the key challenge of detecting small pathological changes in the presence of relatively large and variable time-skew. This is done through an algorithm based on dynamic time-warping and the Lomb-Scargle periodogram.

We evaluated MV on a large population of post-NSTEACS patients from the MERLIN TIMI36 trial. Our results show that high MV is strongly associated with both cardiovascular mortality and sudden cardiac death in patients with preserved LVEF. This holds true even after adjusting for the TIMI risk score. Moreover, for the endpoint of sudden cardiac death, MV is the only long-term ECG metric that is associated with events over follow-up. We are encouraged by these results, since they suggest that MV is not only independent of other clinical measures, but moreover, it may potentially address cases where methods such as echocardiography and other long-term ECG risk variables fail to find patients who could benefit from more aggressive care. We also show how MV can be useful in quantifying the effects of treatments, and potentially troubleshooting new treatments by evaluating their impact on different physiological processes.

In our work on MV, we focus on the micro-level variability in signals, i.e., subtle variability among "normal looking" ECG beats. We ignore the different kinds of beats that occur over time. In the next chapter, we turn our attention towards symbolic analysis, which is a complementary method for studying ECG signals and focuses on the macro-level variability in signals.

Chapter 4

Symbolic Analysis

In this chapter, we present our work on the symbolic analysis of physiological signals. In contrast to the micro-level changes studied by morphologic variability, symbolic analysis focuses on macro-level changes in physiological signals. For ECG, this means that while morphologic variability studies subtle changes within "normal looking" beats, symbolic analysis looks at the different classes of beats that occur over time and ignores subtle changes within these groups.

The central idea underlying our work on symbolic analysis is that symbolization, i.e., the transformation of continuous time series into symbolic sequences (or strings), facilitates many kinds of analysis. In particular, it allows us to represent and search through physiological signals as textual data rather than real-valued time series. This allows us to leverage an extensive body of work on searching and analyzing textual data, including ideas from computational biology and information theory.

In what follows, we first present an overview of symbolic analysis in Section 4.1. We review the concept of symbolization and how it is used in different disciplines, and describe how we can extend this idea to physiological signals. In particular, we discuss why symbolic analysis is an appropriate and powerful paradigm for physiological data, and what domain specific challenges need to be addressed for its use in this setting. We then present an efficient algorithm to symbolize many broad classes of physiological signals in Section 4.2, and demonstrate how symbolization using this algorithm preserves important information in the original signal. After introducing the concept of symbolization and presenting an algorithm to symbolize physiological signals, we turn our attention to the analysis of symbolic signals from multiple patients for risk stratification in Section 4.3. We propose the idea of finding high risk symbolic patterns that are conserved in patients experiencing adverse events, and the idea of finding patients at risk of different adverse outcomes through a comparative approach that groups together patients with similar symbolic sequences (and potentially similar risk profiles). For both these ideas, we propose algorithms that are runtime and space efficient, and can handle very large datasets. Finally, we show how symbolic analysis can be used in settings beyond risk stratification for ACS in Section 4.4.

4.1 Overview of Symbolization and Computational Physiology

4.1.1 Symbolization

Symbolization, or the discretization of raw time series into a sequence of symbols, is widely used in different disciplines (an excellent review on this subject appears in [67]). In many of these applications, the time series of interest cannot be analyzed by traditional tools. This is often due to two factors. The time series may possess complex dynamics that cannot be analyzed by traditional tools. Instead, more sophisticated but computationally expensive analyses are necessary. The runtime of these analyses may be prohibitive. Alternatively, even in cases where simpler analyses may make sense, computational costs are a concern if large amounts of data need to be studied.

In this setting, symbolization can be a powerful tool. It can help reduce the amount of data while retaining much of the important temporal information. An advantage of working with symbols rather than the original time series data is that the efficiency of computations is greatly increased. Furthermore, by abstracting away information that is not relevant, symbolization may enhance our understanding of the process being studied. The analysis of symbolic data is also often less sensitive

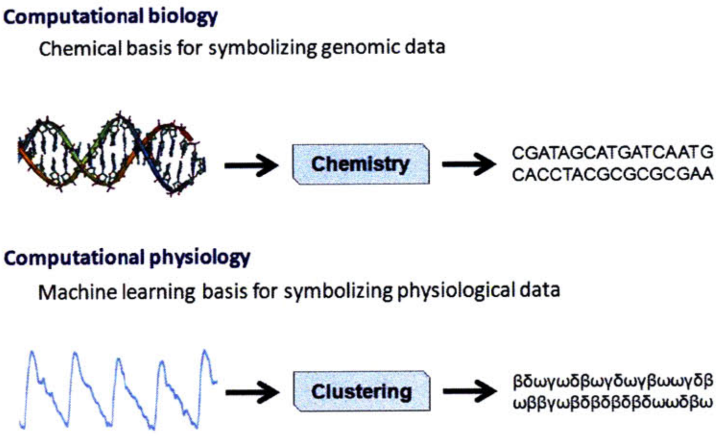


Figure 4-1: Parallel between the use of symbolization in computational biology and in computational physiology.

to measurement noise. Symbolic methods are thus favored in circumstances where complicated analyses are necessary, large amounts of data exists, and where speed, understanding and robustness to noise are important.

4.1.2 Computational Physiology

The factors discussed in Section 4.1.1 are particularly relevant in the study of physiological signals. We therefore approach the analysis of this data within a symbolic framework.

We note that there is a successful precedent for the use of symbolization in the analysis of biological data, i.e., the discipline of computational biology. One of the key early successes of this field was to use a powerful abstraction for representing genomic data as string data. By representing different chemical units as symbols (e.g., guanine = G, cytosine = C), computational biologists were able to transform chemical data into text. This symbolization transformed the data into a form where it was more amenable to analysis, given the extensive literature that exists on the efficient analysis of string data.

We use machine learning (more specifically, clustering methods) in a similar manner to achieve the symbolization of physiological data. This leads to the development of a parallel discipline of *computational physiology* (Figure 4-1).

4.1.3 Challenges

Our concept of computational physiology shares many of the same challenges as computational biology. We study our signals in the absence of significant prior knowledge, and focus instead on using statistics to find interesting patterns and activity. We also deal with the problem of these patterns and activity occurring in an approximate form. While approximate patterns exist in genomic data due to nucleotide mutations, in our work the presence of noise in the original signal, errors associated with symbolization, and randomness in the physiological system generating the signal may all lead to a similar effect.

In addition to this, our work faces some additional challenges. One of the differences between our vision of computational physiology, and the use of symbolization in computational biology, is that in the case of computational biology the symbol definitions are universal and consistent. For example, a cytosine molecular does not change from one patient to another, and as a consequence of this, the mapping cytosine = C does not change either.

For physiological signals, however, the same functional activity may look different across patients (e.g., a heart rate of 95 beats per minute may be normal for one patient, but may be abnormally fast for another). This means that the symbols for different patients are essentially derived from different alphabets, since they cannot be directly compared. Symbolization in the context of physiological signals is therefore associated with the need to relate functionally similar activity across patients. This problem of *symbol registration* may be particularly hard for applications that make use of little or no prior knowledge, and must therefore discover what constitutes the same functional activity across patients as part of the symbolization or subsequent analysis process.

Another challenge associated with the symbolization of physiological signals is determining which segments of the signal should be treated as units for symbolization. For signals with repetitive structure, such as cardiovascular and respiratory data, the quasi-periodic units of activity can be symbolized. However, many important physiological signals are not periodic or quasi-periodic, e.g., EEG data and other neurological signals in general. In these cases, an alternative is necessary, such as segmenting the signal into fixed time windows [121] or stationary segments [122].

Finally, different analyses are necessary while analyzing symbolic representations of physiological data, than for symbolic representations of genomic data. In particular, the problem of risk stratification needs to be framed within the context of symbolic data. This requires the need for new broad areas of analysis and efficient algorithms to solve these problems.

4.1.4 Our Approach

In our work, we propose a two-step process for discovering relevant information in physiological datasets. As a preliminary step, we segment physiological signals into quasi-periodic units (e.g., hearts beats recorded on ECG). These units are then partitioned into classes using morphological features. This allows the original signal to be re-expressed as a symbolic string, corresponding to the sequence of class labels assigned to the underlying units. We also create a prototypical representation of each symbol by aggregating the units that fall into the same class using techniques described in Section 5.2. This allows us to re-represent the original signal as a symbolic sequence, while retaining information about the physiological activity that the specific symbols correspond to.

The second step involves searching for significant patterns in the reduced representation resulting from symbolization. In the absence of prior knowledge, significance is assessed by organization of basic units as adjacent repeats, frequently occurring words, or subsequences that co-occur temporally with activity in other signals. The fundamental idea is to search for variations that are unlikely to occur purely by chance, since such patterns are more likely to be clinically relevant. For the multi-patient case, we propose methods that are robust and allow for symbols to be drawn from different alphabets to be compared.

Figure 4-2 presents an overview of this approach. We start by using conventional techniques to segment an ECG signal into individual beats. The beats are then automatically partitioned into classes based upon their morphological properties. For the data in Figure 4-2, our algorithm found five distinct classes of beats, denoted in the figure by the arbitrary symbols θ, γ, β, α, and ψ (Figure 4-2). For each class an archetypal beat is constructed that provides an easily understood visible representation of the types of beats in that class. The original ECG signal is then replaced by the corresponding sequence of symbols. This process allows us to shift from the analysis of raw signals to the analysis of symbolic strings. The discrete symbolic representation provides a layer of data reduction, reducing the data rate from 3960 bits/second (sampling at 360 Hz with 11 bit quantization) to n bits/second (where n depends upon the number of bits needed to differentiate between symbols, three for this example). Finally, various techniques are used to find segments of the symbol sequence that are of potential clinical interest. In this example, a search for approximate repeating patterns found the rhythm shown in Figure 4-2. The corresponding prototypical representation in Figure 4-2 allows this activity to be readily visualized in a compact form.

This example helps illustrate an important advantage of symbolization. Finding interesting activity similar to that shown in Figure 4-2 would have been hard while operating directly on the time series signal. However, within a symbolic framework, it could be discovered efficiently. We believe that symbolization offers an elegant way to carry out analyses that are otherwise hard for time series signals.

4.2 Creating Symbolic Representations

An extensive literature exists on the subject of symbolization [67]. Essentially, the task of symbolizing data can be divided into two sub-tasks. As a first step, the signal needs to be segmented into intervals of activity. Following this, the set of segments is partitioned into classes and a label associated with each class. The segmentation stage decomposes the continuous input signal into intervals with biologically relevant boundaries. A natural approach to achieve this is to segment the physiological signals

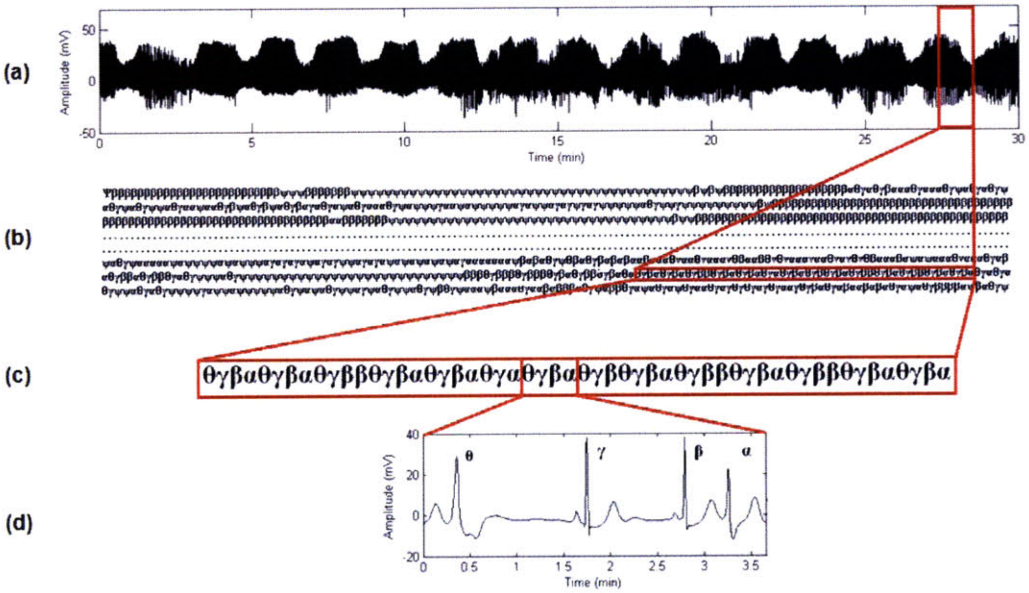


Figure 4-2: Overview of symbolic analysis: (a) Raw data corresponding to Patient 106 in the MIT-BIH Arrhythmia Database. The red rectangle denotes a particular pattern hidden within the raw data. This pattern is difficult to identify by visual examination of the original signal alone. (b) The raw ECG data is mapped into a symbolic representation (11 lines of the symbol sequence are elided from this figure). (c) An example rhythm of a repeating sequence, found in the symbolized representation of a region of data corresponding to the boxed area of the raw data in (a). (d) An archetypal representation, created using the techniques in Section 5.2, of the repeating signal.

according to some well-defined notion. In this work, we use R-R intervals for heart beats and peaks of inspiration and expiration for respiratory cycles. Since most cardiovascular signals are quasi-periodic, we can exploit cyclostationarity for data segmentation [66].

We treat the task of partitioning as a data clustering problem. Roughly speaking, the goal is to partition the set of segments into the smallest number of clusters such that each segment within a cluster represents the same underlying physiological activity. For example, in the case of ECG data, one cluster might contain only ventricular beats (i.e., beats arising from the ventricular cavities in the heart) and another only junctional beats (i.e., beats arising from a region of the heart called the atrioventricular junction). Each of these beats has different morphological characteristics that enable us to place them in different clusters. There is a set of generally accepted labels that cardiologists use to differentiate distinct kinds of heart beats. Although cardiologists frequently disagree about what label should be applied to some beats, labels supplied by cardiologists provide a useful way to check whether or not the beats in a cluster represent the same underlying physiological activity.

In many cases finer distinctions than provided by conventional classification can be clinically relevant. Normal beats, for example, are usually defined as beats that have morphologic characteristics that fall within a relatively broad range; e.g., QRS complex less than 120 ms and PR interval less than 200 ms. Nevertheless, it may be clinically useful to further divide "normal" beats into multiple classes since some normal beats have subtle morphological features that are associated with clinically relevant states. One example of this phenomenon is Wolff-Parkinson-White (WPW) syndrome. In this disorder, patients have ECG beats that appear grossly normal, yet on close inspection, their QRS complexes contain a subtle deflection called a d-wave and a short PR interval [66]. Since such patients are predisposed to arrhythmias, the identification of this electrocardiographic finding is of interest [66]. For reasons such as this, standard labels cannot be used to determine the appropriate number of clusters.

In our work, we avoid the use of significant prior knowledge and instead use the raw samples of each segment as features. Many different clustering algorithms can then be applied for the unsupervised labeling of a collection of individual observations into characteristic classes ([65] provides a detailed examination of a number of methods that have been used to label ECG beats). Unfortunately, most of these methods are computationally intensive and also tend to ignore small clusters (new clusters are created to better fit high density regions of the clustering space, rather than to fit clusters that are different from other groups but comprise a small number of observations). We attempt to address both these shortcomings, and develop clustering methods that are efficient and also have a higher sensitivity (i.e., can discover classes that occur rarely during the course of a recording) than the techniques described in [65].

4.2.1 Max-Min Clustering with Dynamic Time-Warping

We make use of Max-Min clustering to separate segmented units of cardiovascular signals into groups. The partitioning proceeds in a greedy manner, identifying a new group at each iteration that is maximally separated from existing groups.

Central to this clustering process is the method used to measure the distance between two segments. As was described in Section 3.2.3, one of the challenges of comparing the morphology of segments (defined as the raw samples of each segment) is that simple metrics such as the Euclidean distance are insufficient. The presence of time-skew, which is a common occurrence in physiological signals and leads to these signals being variably dilated or shrunk, leads to physiologically different activity being compared when one segment is blindly subtracted from the other. We therefore adopt our earlier approach of using dynamic time-warping (DTW) to align samples in segments before measuring the differences between them (Section 3.2.3).

The use of DTW means that segments cannot be compared on the basis of individual features (i.e., each raw sample), but instead we can only measure how different two segments are. This creates the need for clustering algorithms that are not feature driven, but distance driven. In [62] and [63], clustering methods are proposed that build on top of DTW. A modified fuzzy clustering approach is described in [62], while

[63] explores the use of hierarchical clustering. Denoting the number of observations to be clustered as N, both methods require a total of $O(N^2)$ comparisons to calculate the dissimilarity between every pair of observations. If each observation has length M, the time taken for each dissimilarity comparison is $O(M^2)$. Therefore, the total running time for the clustering methods in [62] and [63] is $O(M^2N^2)$. Additionally, storing the entire matrix of comparisons between every pair of observations requires $O(N^2)$ space. For very large datasets, the runtime and space requirements are prohibitive.

To reduce the requirements in terms of running time and space, we employ Max-Min clustering [64], which can be implemented to discover k clusters using $O(Nk)$ comparisons. This leads to a total running time of $O(M^2Nk)$, with an $O(N)$ space requirement.

Max-Min clustering proceeds by choosing an observation at random as the first centroid c_1 and setting the set S of centroids to $\{c_1\}$. During the i-th iteration, c_i is chosen such that it maximizes the minimum distance between c_i and observations in S:

$$c_i = \arg\max_{x \notin S} \min_{y \in S} C(x, y) \tag{4.1}$$

where $C(x, y)$ is defined as in Equation 3.9. The set S is incremented at the end of each iteration such that $S = S \cup ci$.

The number of clusters discovered by Max-Min clustering is chosen by iterating until the maximized minimum dissimilarity measure in Equation 4.1 falls below a specified threshold θ. Therefore the number of clusters, k, depends on the separability of the underlying data to be clustered.

The running time of $O(M_2Nk)$ can be further reduced by exploiting the fact that in many cases two observations may be sufficiently similar that it is not necessary to calculate the optimal alignment between them. A preliminary processing block that identifies c such homogeneous groups from N observations without alignment of time-samples will reduce the number of DTW comparisons, each of which is $O(M_2)$, from $O(Nk)$ to $O(ck)$. This pre-clustering can be achieved in a computationally inexpensive manner through an initial round of Max-Min clustering using a simple distance metric.

The running time using pre-clustering is given by $O(MNc) + O(M_2ck)$. The asymptotic worst case behavior with this approach is still $O(M_2Nk)$, e.g., when all the observations are sufficiently different that $c = N$. However, for the ECG data we have examined, c is an order of magnitude less than N. For example, pre-clustering with a hierarchical Max-Min approach yielded a speedup factor of 12 on the data from the MIT-BIH Arrhythmia database used in the evaluation presented later in this chapter.

4.2.2 Evaluation

In what follows, we evaluate our work on symbolization by applying it to ECG datasets. We first report on the use of Max-Min clustering with a DTW-based distance metric to partition ECG beats at a finer granularity than existing clinical labels. We then study how symbolization retains useful information in the original signal by reporting the results of simple analyses on symbolized ECG data.

We stress that these studies are intended purely for evaluation. Neither set of experiments addresses a direct goal of our work. Instead, we use these experiments for illustrative purposes, to explore the strengths and weaknesses of our work on symbolization.

ECG Symbolization by Max-Min Clustering with Dynamic Time-Warping

We applied symbolization to electrocardiographic data in the Physionet MIT-BIH Arrhythmia database, which contains excerpts of two-channel ECG sampled at 360 Hz per channel with 11-bit resolution. Activity is hand-annotated by cardiologists, allowing our findings to be validated against human specialists.

For each patient in the database, we searched for different classes of ECG activity between consecutive R waves within each QRS complex. A Max-Min threshold of

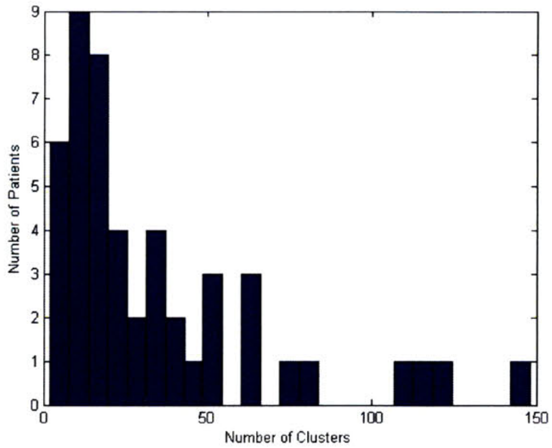


Figure 4-3: Histogram of clusters per patients: The number of clusters determined automatically per patient is distributed as shown, with a median value of 22.

$\theta = 50$ was used, with this value being chosen experimentally to produce a small number of clusters, while generally separating out clinical classes of activity for each patient. As we report at the end of this section, a prospective study on blind data not used during the original design of our algorithm shows that the value of the θ parameter generalizes quite well.

Beats were segmented using the algorithm described in [79]. A histogram for the number of clusters found automatically for each patient is provided in Figure 4-3. For the median patient, 2202 distinct beats were partitioned into 22 classes. A much larger number of clusters were found in some cases, in particular patients 105, 203, 207 and 222. These files are described in the MIT-BIH Arrhythmia database as being difficult to analyze owing to considerable high-grade baseline noise and muscle artifact noise. This leads to highly dissimilar beats, and also makes the ECG signals difficult to segment. For patient 207, the problem is compounded by the presence of multiform premature ventricular contractions (PVCs). Collectively, these records are characterized by long runs of beats corresponding to singleton clusters, which can be easily detected and discarded (i.e., long periods of time where every segmented unit looks significantly different from everything else encountered).

Our algorithm clusters data without incorporating prior, domain-specific knowledge. As such, our method was not designed to solve the classification problem of placing beats into pre-specified clinical classes corresponding to cardiologist labels. Nevertheless, a comparison between our clustering algorithm and cardiologist provided labels is of interest. Therefore we compared our partitioning of the data to cardiologist provided labels included in the MIT-BIH Arrhythmia database. There are a number of ways to compare a clustering produced by our algorithm (CA) to the implicit clustering which is defined by cardiologist supplied labels (CL).

CA and CL are said to be isomorphic if for every pair of beats, the beats are in the same cluster in CA if and only if they are in the same cluster in CL. If CA and CL are isomorphic, our algorithm has duplicated the clustering provided by cardiologists. In most cases CA and CL will not be isomorphic because our algorithm typically produces more clusters than are traditionally defined by cardiologists. We view this as an advantage of our approach since it allow our method to identify new morphologies and patterns that may be of clinical interest.

We say that CA is consistent with CL if an isomorphism between the two can be created by merging clusters in CA. For example, two beats in an ECG data stream may have abnormally long lengths and therefore represent "wide-complex" beats. However, if they have sufficiently different morphologies, they will be placed in different clusters. We can facilitate the creation of an isomorphism between CA and CL by merging all clusters in CA that consist of wide-complex beats. While consistency is a useful property, it is not sufficient. For example, if every cluster in CA contained exactly one beat, it would be consistent with CL. As discussed above, however, in most cases our algorithm produces a reasonable number of clusters.

To determine whether our algorithm generates a clustering that is consistent with cardiologists supplied labels, we examined the labels of beats in each cluster and assigned the cluster a label corresponding to its majority element. For example, a cluster containing 1381 normal beats, and 2 atrial premature beats would be labeled as being normal. Beats in the original signal were then assigned the labels of their clusters (e.g., the 2 atrial beats in the above example would be labeled as normal). Finally, we tabulate the differences between the labels generated by this process and the cardiologist supplied labels in the database. This procedure identifies, and effectively merges, clusters that contain similar types of beats.

We considered only classes of activity that occurred in at least 5% of the patients in the population, i.e., 3 or more patients in the MIT-BIH Arrhythmia database. Specifically, even though we successfully detected the presence of atrial escape beats in patient 223 of the MIT-BIH Arrhythmia database and ventricular escape beats in patient 207, we do not report these results in the subsequent discussion since no other patients in the population had atrial or ventricular escape activity and it is hard to generalize from performance on a single individual. During the evaluation process, labels that occur fewer than three times in the original labeling for a patient (i.e, less than 0.1% of the time) were also ignored.

Tables 4.1 and 4.2 show the result of this testing process. We document differences between the labeling generated by our process and the cardiologist supplied labels appearing in the database. Differences do not necessarily represent errors. Visual inspection of these differences by a board-certified cardiologist, who was not involved in the initial labeling of beats in the Physionet MIT-BIH Arrhythmia database, indicates that experts can disagree on the appropriate labeling of many of the beats where the classification differed. Nevertheless, for simplicity we will henceforth refer to “differences” as “errors.”

In Table 4.1, for the purpose of compactly presenting results, we organize clinical activity into the following groups:

- N = Normal
- Atr = Atrial (atrial premature beats, aberrated atrial premature beats and atrial ectopic beats)
- Ven = Ventricular (premature ventricular contractions, ventricular ectopic beats and fusion of normal and ventricular beats)
- Bbb = Bundle branch block (left and right bundle branch block beats)
- Jct = Junctional (premature junctional beats and junctional escape beats)

Table 4.1: Beats detected for each patient in the MIT-BIT Arrhythmia database using symbolization. To compactly display results we group the clinical classes (Mis = mislabeled beat).

Patient	N	Atr	Ven	Bbb	Jct	Oth	Mis	Mis %
100	2234/2234	30/33					3/2267	0.13%
101	1852/1852	3/3					0/1855	0.00%
102	14/99		4/4			2077/2079	87/2182	3.99%
103	2076/2076						0/2076	0.00%
104	51/163					2027/2040	125/2203	5.67%
105	2530/2534		39/40				5/2574	0.19%
106	1500/1500		508/511				3/2011	0.15%
107			59/59			2074/2075	1/2134	0.05%
108	1748/1748	1/4	17/18				4/1770	0.23%
109			37/40	2486/2486			3/2526	0 12%
111	2117/2117						0/2117	0.00%
112	2533/2533						0/2533	0.00%
113	1782/1782	5/5					0/1787	0.00%
114	1815/1815	4/8	47/48				5/1871	0.27%
115	1946/1946						0/1946	0.00%
116	2281/2281		107/107				0/2388	0.00%
117	1528/1528						0/1528	0.00%
118		82/96	16/16	2147/2161			28/2273	1 23%
119	1540/1540		443/443				0/1983	0.00%
121	1858/1858						0/1858	0.00%
122	2475/2475						0/2475	0.00%
123	1510/1510						0/1510	0.00%
124			52/52	1523/1526	6/34		31/1612	1.92%
200	1737/1739	1/29	796/815				49/2583	1.90%
201	1605/1605	65/76	184/185		3/11		20/1877	1.07%
202	2043/2046	32/48	18/20				21/2114	0.99%
203	2432/2442		318/345				37/2787	1.33%
205	2564/2565	1/3	76/77				4/2645	0.15%
207		114/116	190/208	1538/1559			41/1883	2.18%
208	1507/1575		1327/1348				89/2923	3.04%
209	2603/2617	317/383					80/3000	2.67%
210	2411/2416	14/21	164/183				31/2620	1.18%
212	920/920			1821/1824			3/2744	0.11%
213	2632/2635	4/28	321/581				287/3244	8.85%
214		260/261	1980/1993			14/2254	0.62%	
215	3190/3191		156/159				4/3350	0.12%
217	229/242		138/157			1720/1802	114/2201	5.18%
219	2077/2077	0/7	31/63				39/2147	1.82%
220	1942/1947	91/93					7/2040	0.34%
221	2028/2028		381/382				1/2410	
							0.04%	
222	1939/1977	121/187			125/216		195/2380	8.19%
223	2021/2025	20/89	462/484				95/2598	3.66%
228	1685/1687	0/3	366/371				10/2061	0.49%
230	2249/2249						0/2249	0.00%
231	312/312			1246/1247			1/1559	0.06%
232		1407/1423		435/437			18/1860	0.97%
233	2219/2220	0/7	814/828				22/3055	0.72%
234	2695/2696		3/3		35/50		16/2749	0.58%
Total Beats	76430/768022312/2662		7334/7808	13176/13233	169/311	7898/7996	1493/108812	1.37%
Total Patients	41/41	18/21	29/29	8/8	4/4	4/4		

Table 4.2: Summary comparison of detection through symbolization to cardiologist supplied labels. The labels used correspond to the original MIT-BIH Arrhythmia database annotations (N = normal, L = left bundle branch block, R = right bundle branch block, A = atrial premature beats, a = aberrated atrial premature beats, V = premature ventricular complex, P = paced beat, f = fusion of normal and paced beat, F = fusion of ventricular and normal beat j = junctional escape beat). The top row is indicative of how well the clustering did at identifying the presence of classes of clinical activity identified by the cardiologists for each patient. The bottom row indicates how well the clustering did at assigning individual beats to the same classes as the cardiologists.

	N	L	R	A	a	V	P	f	F	j
Percentage of total patients detected	100.0	100.0	100.0	84.21	100.0	100.0	100.0	100.0	75.0	100.0
Percentage of total beats detected	99.52	99.50	99.67	87.30	85.11	96.80	99.91	78.75	46.69	56.96

- Oth = Others

The overall misclassification percentage is approximately 1.4%.

Figure 4-4 also illustrates how the mislabeling error associated with our clustering approach is distributed across patients. In the majority of the patients, there is less than 1% error.

The data in the first row of Table 4.2 sheds light on critical errors; i.e. errors that cause one to conclude that a patient does not exhibit a certain type of beat when in fact, their ECG signal does contain a significant number of the beats in question. More precisely, we say that a critical error has occurred when a patient has at least three instances of a clinically relevant type of beat and there does not exist at least one cluster in which that beat is a majority element. For example, for each patient for whom the cardiologists found three or more "premature ventricular complexes," the algorithm formed a cluster for beats of that type. On the other hand, for one quarter of the patients with at least three "fusion of ventricular and normal beats," the algorithm did not form a cluster for that type of beat. In 43 out of 48 patients

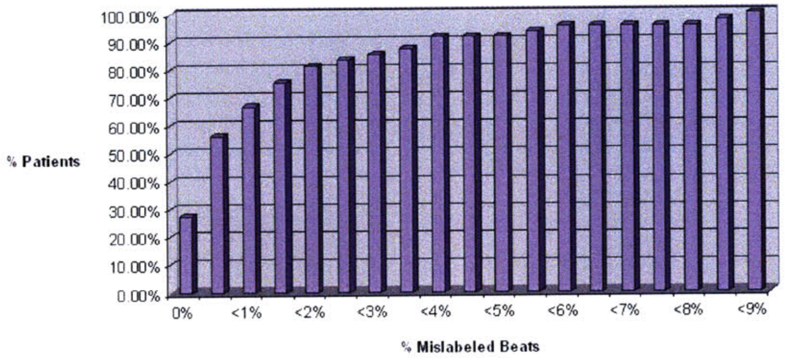


Figure 4-4: Mislabeling Error: Over a quarter of the patients had no mislabeling errors using our clustering approach, over 65% had less than 1% mislabeled beats relative to cardiologist labels.

there were no critical errors. This is important because, in the presence of critical errors, an inspection of the data through visualization of the cluster representatives would conceal the presence of some activity in the dataset. Avoiding critical errors is a challenge because for some patients, the number of elements in different clinical classes varies by a few orders of magnitude.

As Tables 4.1 and 4.2 indicate our symbolization technique does a reasonably good job both at identifying clinically relevant clusters and at assigning individual beats to the appropriate cluster.

For some classes of activity, however, our morphology-based clustering generated labels different from those provided by the cardiologists. Figure 4-5 presents an example where morphology-based clustering differed from the labels in the database. However, given the similarity between the beats labeled F and N in the database, it is not clear that our algorithm is in error. Similarly, our algorithm also failed to distinguish right bundle branch block and junctional premature beats, as shown in Figure 4-6.

Sometimes our algorithm places beats for which cardiologists have supplied the same label into different clusters. As was discussed above, this is not necessarily a bad thing as subtle distinctions between "normal" beats may contain useful clinical information. Figures 4-7 and 4-8 present instances in which our algorithm separated beats

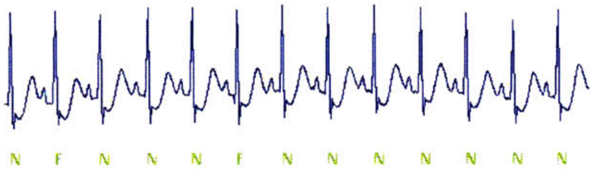


Figure 4-5: Raw tracing of ECG for patient 213 in the MIT-BIH database with fusion of ventricular and normal beats: A sequence of ECG is shown containing beats labeled as both normal (N) and fusion (F). The morphological differences between the two classes of beats are subtle. This excerpt corresponds to time 4:15 in the recording.

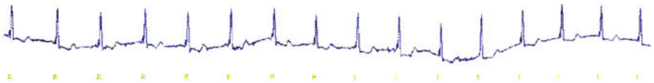


Figure 4-6: Raw tracing of ECG for patient 124 in the MIT-BIH database with junctional escape beats: A sequence of ECG is shown containing both right bundle branch block (R) and junctional premature (J) beats. The morphological differences between the two classes of beats are again subtle. This excerpt corresponds to time 4:39 in the recording.

that were assigned the same label by cardiologists. In Figure 4-7, morphology-based analysis is able to distinguish changes in length. In Figure 4-8, changes in amplitude are discerned automatically. These morphological differences may represent clinically important distinctions. In each instance, beats that are classified as "normal" have very different morphologic features that may be associated with important disease states. Abrupt changes in the R-R interval, like that noted in Figure 4-7, correspond to rapid fluctuations in the heart rate; a finding which can be associated with a number of clinically important conditions such as Sick Sinus Syndrome (SSS) or sinus arrhythmia [66]. Similarly, significant changes in QRS amplitude, like that seen in Figure 4-8, can be observed in patients with large pericardial effusions [66]. Both of these diagnoses are important syndromes that can be associated with adverse clinical outcomes. Therefore we view the ability to make such distinctions between beats as a benefit of the method.

Data from the MIT-BIH Arrhythmia database were used during the initial design of the symbolization algorithm, and the results reported in Tables 4.1 and 4.2 were

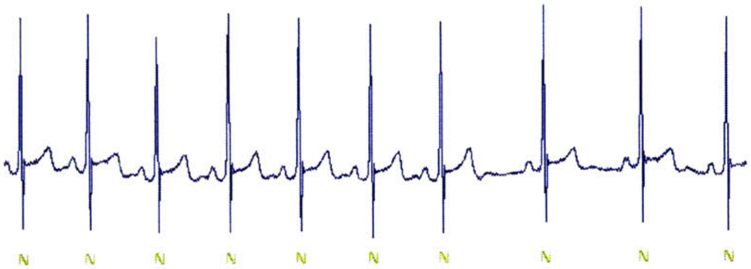


Figure 4-7: Raw tracing of ECG for patient 115 in the MIT-BIH database with normal beats: A sequence of ECG is shown containing normal beats. This sequence represents an example where morphology-based analysis separates the beats into short (first 7 beats) and long (last three beats) classes. The beats still fall in the same clinical class, but this separation, which indicates an abrupt change in heart rate, may potentially be of interest for the purpose of higher level analysis. This excerpt corresponds to time 7:40 in the recording.

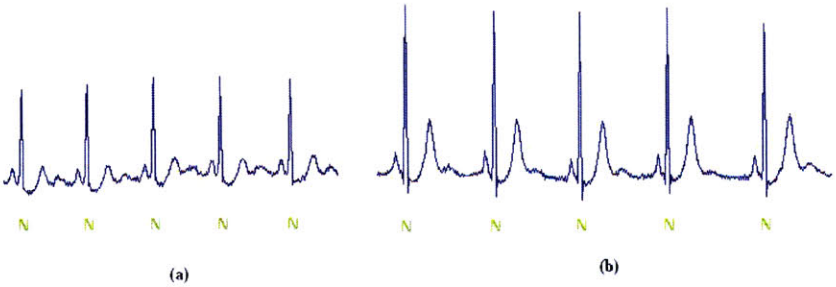


Figure 4-8: Raw tracing of ECG for patient 106 in the MIT-BIH database with normal beats: (a) ECG corresponding to time 16:54 in the file. (b) ECG corresponding to time 21:26 in the file. Morphology-based analysis places the beats shown in (a) and (b) into separate clusters based on changes in amplitude.

Table 4.3: Summary comparison of detection through symbolization to cardiologist supplied labels for the MGH/MF Waveform database. The labels of the columns match those in Table 4.2 with J = junctional premature beats.

	N	V	P	J	F
Percentage of total patients detected	100.0	100.0	100.0	100.0	100.0
Percentage of total beats detected	99.91	96.51	98.84	100.0	100.0

generated on this data set. To test the robustness of the method, we also tested our algorithm on ECG data on the first forty patients from the MGH/MF Waveform database (i.e., mgh001-mgh040), which was not used in design of the algorithm. This dataset contains fewer episodes of interesting arrhythmic activity than the MIT-BIH Arrhythmia database and is also relatively noisy, but contains ECG signals sampled at the same rate (i.e., 360 Hz) with 12 bit resolution; i.e., a sampling rate and resolution similar to that of the MIT-BIH Arrhythmia database. The recordings are also typically an hour long instead of 30 minutes for the MIT-BIH Arrhythmia database.

Table 4.3 shows the performance of the symbolization algorithm on this dataset. The results are comparable to the ones obtained for the MIT-BIH Arrhythmia dataset. The median number of clusters found in this case was 43. We removed file mgh026 from analysis because of the many errors in the annotation file which prevented any meaningful comparisons against the cardiologist provided labels. We also removed file mgh002, which was corrupted by noise that led to errors in the segmentation of the ECG signal. We also detected the presence of atrial escape beats for patient mgh018, but do not report results for this class in Table 4.3 since no other patients revealed similar activity.

Preservation of Information in Symbolized Data

We supplement our evaluation of Max-Min clustering with a DTW distance metric by studying how symbolization retains useful information in the original signal. We do this by reporting the results of simple analyses on symbolized ECG data. We

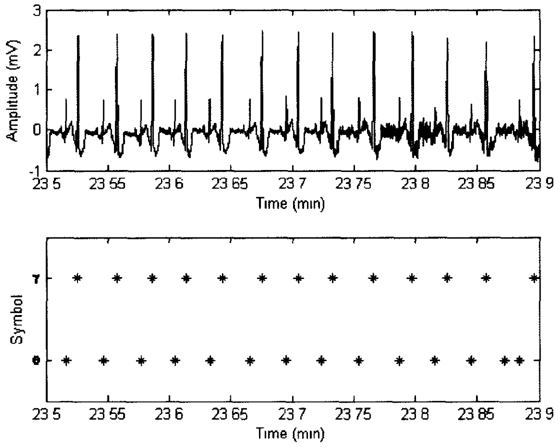


Figure 4-9: A patient with ventricular bigeminy.

show that even with these simple analyses, important information related to many pathological conditions is preserved in the symbolized ECG signal.

Figures 4-9 and 4-10 provide examples of applying techniques to detect approximate tandem repeats. The figures show a fragment of the raw signal and a pictorial representation of the symbol stream for that fragment. The pictorial representation provides a compact display of the symbol string and facilitates viewing the signal over long time intervals. In each case, the repeating sequence in the symbolic signal corresponds to a well-known cardiac rhythm that can be recognized in the raw tracings. Figure 4-9 presents a signal showing a ventricular bigeminy pattern, while Figure 4-10 shows trigeminy. The associated symbolic streams provided for both figures show the repetitious activity in the reduced symbolic representations.

Figure 4-11 shows how searching for approximate tandem repeats in symbolized data can discover complex rhythms that are easy for clinicians to miss. In this case, approximate repeat detection identifies an intricate pattern which likely represents episodes of an ectopic atrial rhythm with aberrant ventricular conduction superimposed on an underlying sinus rhythm. This clinically significant rhythm was not marked by the clinicians who annotated the signal.

Figure 4-12 shows an example in which the detection of recurrent transient patterns in symbolic signals reveals many short, unsustained episodes of tachyarrhythmic

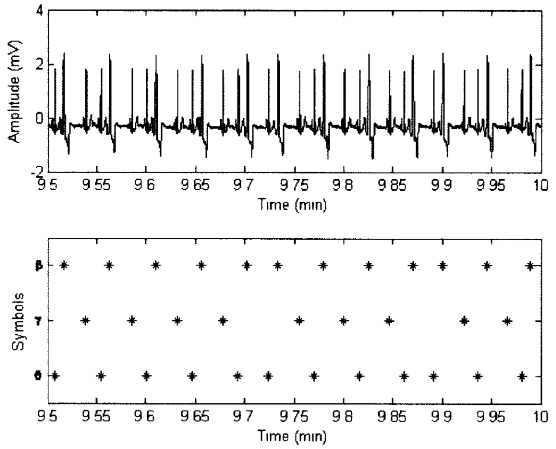


Figure 4-10: A patient with ventricular trigeminy.

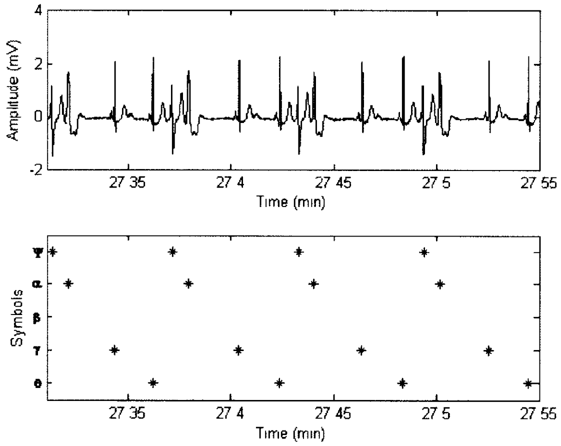


Figure 4-11: A rhythm of 4 units corresponding to an ectopic atrial rhythm.

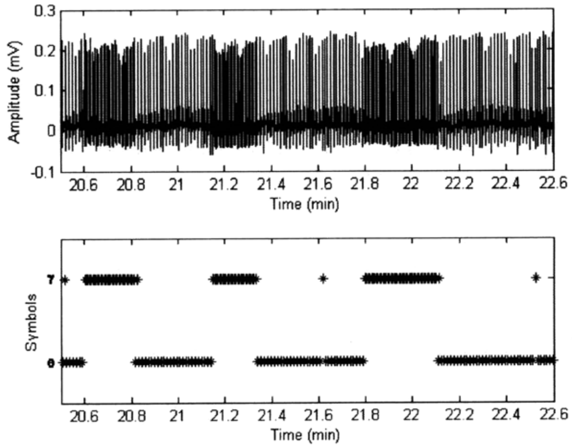


Figure 4-12: A patient with recurrent tachyarrhythmic episodes. These episodes appear in the raw tracing as dense regions, corresponding to an increased number of heart beats during these periods owing to faster heart rate.

activity. The tachyarrhythmic beats occur infrequently relative to normal beats, and consecutive runs of such activity are unlikely to have occurred merely at random.

Figure 4-13 presents the result of applying techniques to discover high entropy segments. These techniques are able to discover segments of ECG corresponding to atrial fibrillation. The irregularity of activity leads to entropy increasing noticeably in windows of the symbolic stream, owing to the unstructured nature of the underlying disorder.

Figures 4-14 and 4-15 demonstrate multi-signal trend detection on symbolized data. In Figure 4-14 the search for correlated activity revealed a case of pulsus paradoxus, where inspiration is associated with a significant drop in arterial blood pressure. This is often associated with cardiac tamponade, severe COPD, pulmonary embolism or right ventricular infarction. In Figure 4-15 episodes of faster heart rate can be seen to occur in conjunction with increased arterial blood pressure, a finding indicative of a hemodynamically significant rhythm. In both cases, associations between the symbolic representations allow for these phenomena to be easily detected.

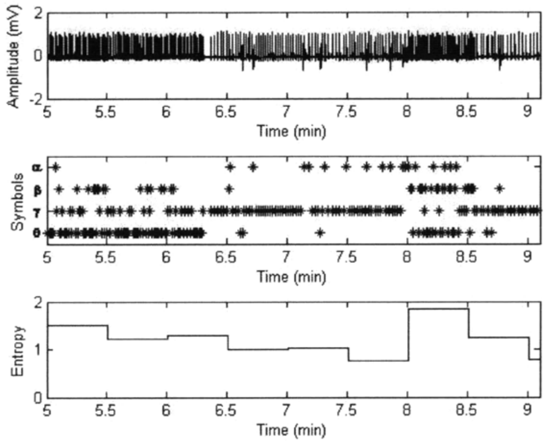


Figure 4-13: Raw ECG tracing, symbolic signal and entropy taken over 30 second windows for a patient with atrial fibrillation. As in Figure 14, atrial fibrillation in the raw tracings corresponds to the dense regions.

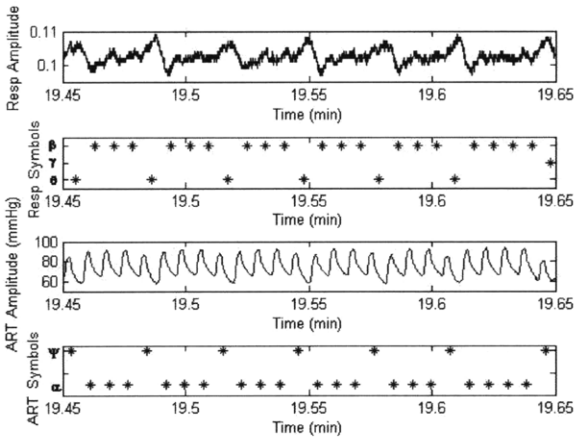


Figure 4-14: Respiration and arterial blood pressure signals for a patient with pulsus paradoxus.

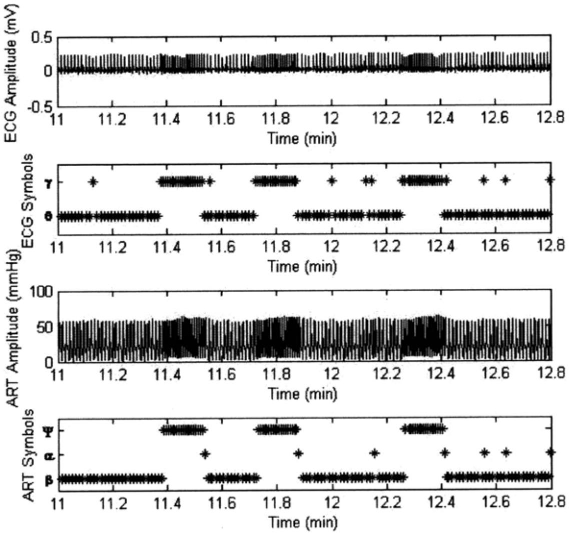


Figure 4-15: ECG and arterial blood pressure signals for a patient in whom fast heart rate leads to increased arterial blood pressure.

4.3 Symbolic Analysis

In the remainder of this section, we focus our attention on analyzing symbolic sequences from multiple patients with the goal of risk stratification. We approach risk stratification in two different ways.

A first approach is to study symbolic sequences from multiple patients to find activity that has a statistically significant presence or absence in sub-populations, i.e., we find sub-sequences that are consistently present or absent in the symbolic sequences for patients who have events. In what follows, we describe the discovery of these patterns only using data from patients who experience events (Section 4.3.1) and using data both from patients who have and do not have events (Section 4.3.3).

A second approach for risk stratification within a symbolic framework is to identify patients with similar symbol dynamics over long periods of time. The hypothesis underlying this work is that patients with similar symbolic signals will have similar risk profiles. We develop this proposed idea further, and describe algorithms for carrying out this analysis in Section 4.3.5.

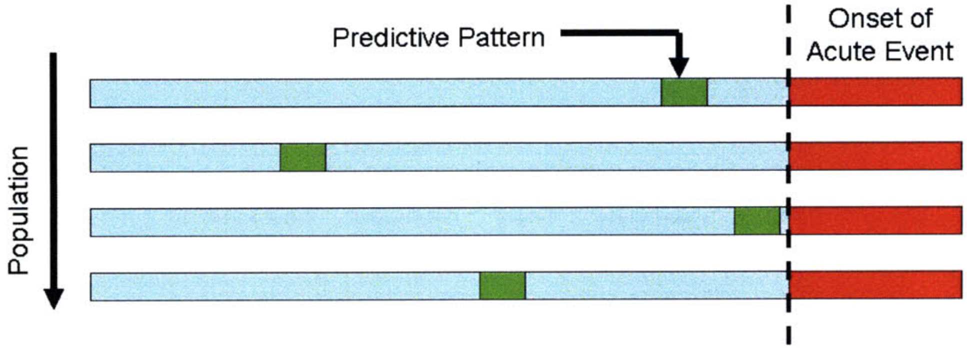


Figure 4-16: Prediction through conservation in the context of a population of patients affected by a common acute clinical event

4.3.1 Pattern Discovery in Positive Examples

We model prediction as the problem of identifying activity that consistently precedes an event of interest. In the absence of any prior knowledge, this activity can be discovered by observing multiple occurrences of the event and detecting statistically significant commonalities in the data preceding it, i.e., by searching for conserved elements unlikely to occur purely by chance prior to the event of interest (Figure 4-16). To handle noise, we further adopt a relaxed view of conservation, whereby precursors may approximately match or be altogether absent on some observations of the event. A further practical consideration is that the search be computationally efficient to handle large amounts of data resulting from multiple observations.

This model of prediction is similar to the search for regulatory motifs in the setting of computational biology. Motif discovery techniques operate on genomic datasets and search for DNA sequences that are conserved across genomes. We generalize this model and describe how the search for precursors to acute clinical events can be carried out in an analogous manner, by first converting continuous physiological signals into an alphabetical representation, and then mining this representation for conserved activity. A variety of randomized greedy algorithms can be used to efficiently carry out the search for such patterns. We use techniques such as TCM and Gibbs sampling as the foundation of our work, and enhance them to operate on data with highly divergent background distributions of symbols, frequent noise and patterns of increased degeneracy relative to genomic data.

In what follows, we describe the proposed unsupervised inference methodology. While the techniques we suggest can be used on a variety of signals and are sufficiently general-purpose, we motivate them in the more concrete setting of searching for predictive activity in physiological signals. We detail the challenges associated with such an approach and describe its benefits and limitations.

Creating Symbolic Representations for Multiple Patients

From a knowledge discovery goal, it is appealing to derive the alphabet for symbolization directly from the data itself. Techniques such as those in Section 4.2 can be employed to achieve this goal. While the approach of generating a patient-specific symbolic representation is powerful in its ability to capture significant changes across a patient, it poses the problem that the clusters are derived separately for each patient. This restricts comparisons across a population. A possible means for addressing this issue is to use a semi-supervised approach where the symbols derived for each patient are related by a human expert. This allows for the symbols to be dynamically derived based on characteristics inherent in the data itself, and for these symbols to be related and compared across a population.

For our work on pattern discovery, registering patient-specific symbols obtained by the techniques described in Section 4.2 represents an area of continuing work. The discussion that follows focuses instead on the use of clinical annotations (or of semi-supervised symbols related manually across patients) despite the possible benefits of patient-specific symbols.

Physiological Motifs

In the setting of computational biology, regulatory motifs correspond to short DNA sequences that regulate gene expression. This notion of a genetic switch that controls activity further downstream is well-suited to our model for prediction. We generalize this idea and choose to model regulatory motifs as sequential triggers that precede abrupt clinical events and are conserved across a population of patients owing to an association with the event.

A recent strategy for regulatory motif discovery that has gained popularity is to make use of comparative genomics [52]. This allows for the discovery of regulatory elements by exploiting their evolutionary conservation across related species. Under this approach, regulatory motif discovery can be viewed computationally as finding sequences that are recurrent is a group of strings, upstream of specified endpoints.

The problem of regulatory motif discovery can be stated more formally in either a combinatorial or probabilistic framework [51]. While the two frameworks both attempt to identify similar preceding subsequences, they may lead to slightly different results and require distinct algorithmic techniques.

Combinatorial: Given a set of sequences $\{s_1, ..., s_N\}$ *find a subsequence* $m_1, ..., m_W$ *that occurs in all* s_i *with* k *or fewer differences.*

Probabilistic: Given a set of sequences $\{s_1, ..., s_N\}$ *find a set of starting positions* $\{p_1, ..., p_N\}$ *in the sequences that lead to the best (as defined below) A x W profile matrix M (where A is the number of different symbols in the data and W is the length of the motif).*

For the probabilistic case, the profile matrix is derived from the subsequences of length W immediately following the starting positions $p_1, ..., p_N$ in each of $s_1, ..., s_N$. These subsequences are lined up and the probability of each of the A unique symbols at every one of the W motif positions is estimated. $M(x, y)$ then gives the probability that the motif has character x at position y. The resulting profile matrix can be scored using different criteria with the implicit goal of seeking a non-trivial profile that is strongly conserved at each position and best explains the data. The scoring function most often used is the log-odds likelihood, i.e.:

$$score = \sum_{i=1}^{N} \sum_{j=1}^{W} \log[\frac{M(s_i(p_i + j - 1), j)}{B(s_i(p_i + j - 1))}] \tag{4.2}$$

where B gives the background distribution of each unique symbol in the data. Effectively, this calculates the log-likelihood of a motif while compensating for trivial occurrences that would be seen in the data merely due to the frequent occurrence of certain symbols.

A complication arising in the context of physiological signals is that of the sparsity of abnormal activity. Periods with interesting events are typically separated by long, variable-sized runs of normal behavior, i.e., the distribution of the symbols is significantly skewed in favor of normal labels. This increases the number of trivial motifs in the data and consequently the running time of the motif discovery algorithms. In addition, for algorithms such as TCM and Gibbs sampling discussed shortly, a secondary effect resulting from the presence of long stretches of normal behavior is that the starting locations chosen randomly may often correspond to uninteresting regions of the signal, further increasing time to convergence.

The issue of degeneracy is frequently encountered in DNA sequences and assumes a critical role for physiological motifs as well. Predictive patterns may be approximately conserved across some patients in a population, while in others, they may be missing altogether. This results from a variety of factors, including differences in the age, gender, clinical history, medications and lifestyle of patients, as well as noise obscuring predictive patterns in some recordings.

The goal of detecting imperfectly conserved activity represents a significant challenge to the task of discovering precursors. Since patterns can vary, the process of determining whether a pattern appears in a patient is required to explore a larger search space, spanning all possible variations. Similarly, the fact that some patients may have the predictive activity obscured due to noise requires recognizing these cases and preventing motif discovery algorithms from forcibly incorporating this data in the search process.

Computational Biology Algorithms for Motif Discovery

We review three popular algorithms for finding regulatory motifs using comparative genomics; the Two Component Mixture (TCM) algorithm using expectation-maximization, Gibbs sampling, and Consensus. TCM and Gibbs sampling attempt to solve the probabilistic formulation of motif discovery, while Consensus focuses on the combinatorial problem.

Two Component Mixture (TCM): TCM is an enhancement to the basic EM algorithm [53], which essentially reduces the search into two smaller, decoupled problems. The first (i.e., the M-step) involves constructing the profile for a motif given a set of fuzzy starting positions $p_1, ..., p_N$ in the input sequences (the M-step). The second (i.e., the E-step) then uses this matrix profile representation to score all possible starting positions in every sequence and update the initial $p_1, ..., p_N$.

The overall TCM algorithm operates in the following manner:

```
TCM($\{s_1, ..., s_N\}$, $W$):
1. Set random starting positions $p_1, ..., p_N$
2. Do
    i. M-step to update profile matrix
    ii. E-step to update starting positions
    Until the change in the score of $M$ is less than some threshold $\epsilon$
```

The M-step of TCM estimates the profile matrix using the probability Z_{ij} that the motif starts in sequence i at position j. As a first step, the values $n_{c,k}$ are estimated, which indicate how often the character c occurs at position k in the motif.

$$n_{c,k} = \begin{cases} \sum_i \sum_{j|s_{i,j}=c} Z_{ij} \; k > 0 \\ n_c - \sum_{j=1} W n_{c,j} \; k = 0 \end{cases} \tag{4.3}$$

$k = 0$ represents the case where character c occurs in the sequence outside the motif while n_c gives the total number of times c occurs in the data. Using these values, we can obtain a profile matrix M as follows:

$$M_{c,k} = \frac{n_{c,k} + d_{c,k}}{\sum_a (n_{a,k} + d_{a,k})} \tag{4.4}$$

where $d_{c,k}$ denotes the pseudocount for character c and helps ensure that the probability of c at position k is not zero while estimating frequencies from finite data [53].

In addition to computing the profile matrix during the M-step, TCM also calculates a prior probability that a motif might start arbitrarily at any position in the data. This is denoted by λ and is obtained by taking the average of Z_{ij} across all sequences and positions.

TCM primarily differs from other EM approaches to motif discovery in its E-step. For every sequence si in the dataset TCM assigns a likelihood L_{ij} to the W-mer starting at each position j:

$$L_{ij}(1) = Pr(s_{ij}|Z_{ij} = 1, M, b) = \prod_{k=j}^{j+W-1} M_{k-j+1,c_k} \tag{4.5}$$

and:

$$L_{ij}(0) = Pr(s_{ij}|Z_{ij} = 0, M, b) = \prod_{k=j}^{j+W-1} b_{c_k} \tag{4.6}$$

where b gives the background probability for each character in the dataset. For iteration t of TCM, the values of Z_{ij} can then be estimated using:

$$Z_{ij}^{(t)} = \frac{L_{ij}^{(t)}(1)\lambda^{(t)}}{L_{ij}^{(t)}(0)[1-\lambda^{(t)}] + L_{ij}^{(t)}(1)\lambda^{(t)}} \tag{4.7}$$

Gibbs Sampling: Gibbs sampling [49] can be viewed as a stochastic analogue of EM for finding regulatory motifs and is less susceptible to local minima than EM. It is also much faster and uses less memory in practice. This is because unlike EM, the Gibbs sampling approach keeps track only of the starting locations $p_1, ..., p_N$ of the motif in each sequence and does not maintain a distribution over all possible starting positions for the motif (i.e., the Z_{ij} in TCM representing fuzzy starting positions are replaced by hard $p_1, ..., p_N$).

The Gibbs sampling algorithm for motif discovery can then be written as:

```
GIBBS({$s_1, ..., s_N$}, $W$):
1. Set random initial values for $p$
```

```
2. Do
    i. Select sequence $s_i$ at random
    ii. Estimate $M$ from set $s_1, ..., s_N - s_i$
    iii. Use $M$ to score all starts in $s_i$
    iv. Pick start $p_i$ with probability proportional to its score
    Until the change in the score of $M$ is less than some threshold $\epsilon$
```

Gibbs sampling is less dependent on the initial parameters than TCM and therefore more versatile. However, it is dependent on all sequences having the motif. This is an inefficiency we address in our work.

Consensus: Consensus [50] is a greedy motif clustering algorithm that picks out two sequences at random, finds the most conserved pairs amongst them and then iterates over all the remaining sequences adding the W-mers that match best to the results of the previous iteration at every stage.

The Consensus algorithm is as follows:

```
CONSENSUS($\{s_1, ..., s_N\}$, $W$):
1. Pick sequences $s_i$ and $s_j$ at random
2. Find most similar $W$-mers in $s_i$ and $s_j$
3. For each unprocessed sequence $s_k$
    i. Expand solution set with $W$-mers from $s_k$ that
           match best with previous ones
```

Data Transformations and Subset Based Tecniques

Active Regions: The issue of skewed symbol distributions can be addressed by removing long stretches of activity that are known to be uninteresting. By definition, a predictive motif is associated with an acute clinical event and must be associated with abnormal activity. As a result, trivial motifs comprising normal activity can be trimmed away to reduce the running time associated with the motif-discovery algorithms. For example, given the sequence:

```
V J V J J N N N N N N N N N N N V N V N B B r
```

A possible reduction of this data would be:

```
V J V J J N+ V N+ V N+ B B r
```

This technique is associated with a significant loss of information. Specifically, the search for motifs proceeds in the transformed space, and the N+ regular expression may occur in motifs without a consistent meaning (i.e., it may be arbitrarily long in some patients). The more general issue here is that conservation of a pattern in the transformed space does not imply conservation in the original signals.

To avoid this issue, we identify regions of abnormal activity, i.e., active regions, by splicing out trivial periods in the signal. Given a motif length W, this involves iterating over the data and removing all normal symbols that would occur only in trivial motifs. This approach preserves the temporal structure of abnormal stretches of the signal, ensuring that the motifs correspond to patterns that are conserved in all of the original signals. For example, using this approach for a motif of length 3, the original example pattern above would map to:

```
V J V J J N N V N V N B B r
```

Gibbs2 and Seeded Consensus: The Gibbs sampling algorithm presented earlier assumes that a motif is present in all sequences. To deal with the issue of degeneracy, where noise may obscure the predictive pattern completely for some patients, we propose a new algorithm that provides a layer of robustness while dealing with a population where activity may be altogether absent in some of the observed examples. This is achieved by adding a second layer of Gibbs sampling to the original algorithm, leading to the Gibbs2 algorithm presented here.

The Gibbs2 algorithm operates at any time on a working subset $V = v_1, ..., v_C$ of the original sequences $s_1, ..., s_N$. Sequences are dynamically swapped into and out of this set with the goal of replacing poor matches with potentially better options. The underlying goal is to arrive at a cluster of sequences that share a strongly conserved motif.

The initial subset of sequences is chosen at random, and at each iteration, a single sequence v_i in the working set is scored at every position p_i using the profile generated from $V - v_i$, i.e.:

$$score(v_i(p_i)) = \sum_{j=1} W \log[\frac{M(s_i(p_i + j - 1), j)}{B(s_i(p_i + j - 1))}] \tag{4.8}$$

With some probability v_i is swapped out and replaced by one of the sequences outside the working set. The probability of being swapped out varies inversely with the maximum score seen for the sequence at any position, i.e., the score at the position that corresponds most strongly to the profile matrix:

$$\log[Pr(swap)] \propto -\max_{p_i}[score(v_i(p_i))] \tag{4.9}$$

with the proportionality factor depending on the length of the motifs being searched for.

The intuition behind the Gibbs2 algorithm is that if a sequence scores high for a motif, it matches quite well with other sequences used to derive the profile and is retained with a higher probability. Conversely, if a sequence does not score highly, it matches poorly with the remaining sequences in the working set used to derive the profile.

Ideally, the sequence swapped out should be replaced by one that scores highest on the profile matrix being used. This approach is computationally intensive since all outstanding sequences need to be scored before the optimal one can be chosen. To avoid this, once a sequence is swapped out, it is replaced by any of the sequences outside the working set at random. This avoids the need to score all previously excluded sequences to find the one with the best match. Furthermore, after each swap, further swapping is temporarily disabled to allow the new sequence to be absorbed and contribute to the profile matrix.

The Gibbs2 algorithm can be written as follows (with C denoting the size of the working set and K representing the number of iterations swapping is disabled after a sequence is replaced from one outside the working set):

```
GIBBS($\{s_1, ..., s_N\}$, $W$, $C$, $K$):
1. Choose $C$ sequences at random from $\{s_1, ..., s_N\}$
2. Set random initial values for $p$
3. Do
    i. Select sequence $v_i$ at random
    ii. Estimate $M$ from set $V - v_i$
    iii. Use $M$ to score all starts in $v_i$
    iv. Swap out $v_i$ with $Pr(swap)$ and replace it with a random
          sequence outside the working set
    v. If swap occurs
      a. Disable swapping for $K$ iterations
    vi. Pick start $p_i$ with probability proportional to its score
Until the change in the score of $M$ is less than some threshold $\epsilon$
```

The Gibbs2 approach can be used to iteratively partition the data into a set containing a strongly conserved motif and an outstanding set that can be broken into further subsets sharing a common pattern. This allows for the discovery of multiple predictive motifs occurring in subsets of the population.

We propose choosing the working set size by studying how the log-odds likelihood of motifs changes for different selections of C. The average contribution to the log-odds likelihood by each sequence in the working set can be measured as:

$$\frac{1}{C}\sum_{i=1}^{C}\sum_{j=1}^{W}\log[\frac{M(s_i(p_i+j-1),j)}{B(s_i(p_i+j-1))}] \tag{4.10}$$

As sequences are added to the working set, the average contribution measured decreases significantly if the addition of a further sequence prevents the working set from sharing a common motif, i.e., if the additional sequence does not allow a strong motif to be identified. The size of the working set for the Gibbs2 algorithm can therefore be determined by searching for a knee in the curve relating the average contribution to the log-odds likelihood by each sequence with C. This process may be approximated by a binary search to reduce computation.

The use of Gibbs2 also allows for the performance of the Consensus algorithm to be improved. Specifically, Consensus can be seeded using a strongly conserved pattern obtained by Gibbs2. This reduces the likelihood that Consensus will be affected by a poor choice of the initial two strings.

4.3.2 Evaluation of Pattern Discovery in Positive Examples

We applied our techniques to the Physionet Sudden Cardiac Death Holter Database (SDDB). This database contains several hours of ECG data recorded using Holter monitors from 23 patients who experienced sudden cardiac death. The recordings were obtained in the 1980s in Boston area hospitals and were compiled as part of a later study of ventricular arrhythmias. Owing to the retrospective nature of this collection, there are important limitations. Patient information is limited, and sometimes completely unavailable, including drug regimens and dosages. Furthermore, sudden cardiac death may result from a variety of underlying causes and it is likely that among the 23 patients there are multiple groups sharing different regulatory factors. Despite these shortcomings, the SDDB ECG signals represent an interesting dataset since they represent a population sharing a common acute event. In addition, the recordings are sufficiently long (up to 24 hours prior to death in some cases) that it is likely the predictive factors occurred during the recording period. Finally, the signals in SDDB are generally well-annotated, with cardiologists providing labels at the level of each beat, and this yields a source of clinically relevant symbols that can be used to search for motifs.

For the 23 SDDB patients TCM, Gibbs sampling, Gibbs2 and Consensus were used to discover potentially predictive motifs of lengths 4, 10 and 16. Since TCM, Consensus and the variants of the Gibbs sampling algorithms are stochastic in nature, a hundred runs were executed with the strongest motifs being automatically returned as the solution. The scoring function used was the log-likelihood score.

In each case, the endpoint used to signify the acute event associated with death was the occurrence of ventricular fibrillation (VF). This was annotated for all patients and only regions preceding VF were searched for conserved motifs.

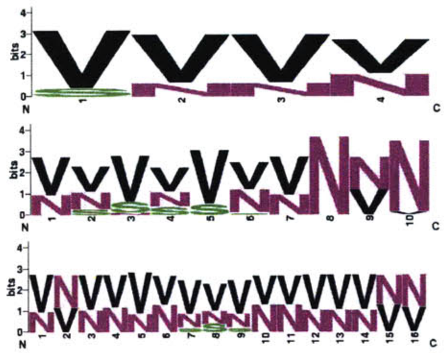


Figure 4-17: Motifs of length 4, 10 and 16 found using TCM.

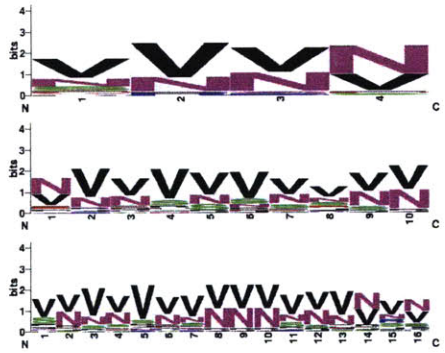


Figure 4-18: Motifs of length 4, 10 and 16 found using Gibbs sampling.

For visualization purposes, we used WebLogo [22] to display the motifs returned by our algorithms. This uses the profile matrix to represent motifs as sequence logos, which are graphical representations consisting of stacks of symbols. For each position in the motif, the overall height of the stack indicates how strongly the motif is conserved at that position, while the height of symbols within the stack indicates the relative frequency of each symbol at that position. For example, for the length 10 motif in Figure 4-17, the sequence logo shows that the motif is strongly conserved at positions 8 and 10, where the predictive sequence was found to contain normal beats across patients. The motif is also conserved at positions 1, 3 and 5, where ventricular activity was seen for most patients, with some occurrences of normal beats (position 1) and supraventricular beats (positions 3 and 5) as well.

For position j in the motif, the height of symbol i at that location is given by:

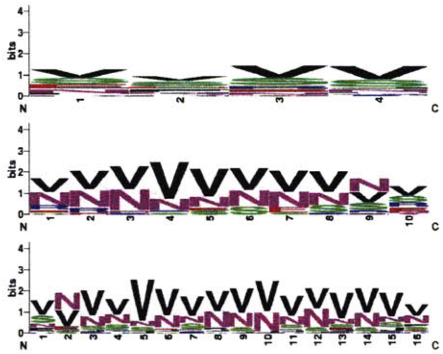


Figure 4-19: Motifs of length 4, 10 and 16 found using Consensus.

$$M(i,j)[2 - H(j)] \tag{4.11}$$

where:

$$H(j) = -\sum_{k} M(k,j)log_2(M(k,j)) \tag{4.12}$$

For Consensus, where a profile matrix is not explicitly constructed, the best-matching subsequences were used to derive a profile that could be represented using WebLogo. This allowed for results to be consistently visualized, irrespective of the algorithm used to discover motifs.

More information on WebLogo can be found at http://weblogo.berkeley.edu.

Data Transformation

The transformations discussed earlier can be evaluated in terms of the data compression realized using these approaches. This allows for an appreciation of the extent to which the original data contains long runs of normal activity that can be compacted. The original sequences across the 23 patients contained 1,216,435 symbols in total, each corresponding to a single beat annotated by a skilled cardiologist. Using the notion of active regions and stripping away uninteresting normal motifs reduced the size of the data to 257,479 characters, i.e., a reduction of 78.83%.

TCM, Gibbs Sampling and Consensus

Figures 4-17 to 4-19 present the results returned by TCM, Gibbs sampling and Consensus as sequence logos. Commonly occurring labels are N=normal, V=premature ventricular contraction, and S=supraventricular premature or ectopic beats.

The motifs discovered by all three algorithms were similar and comprised runs of premature ventricular contractions. For each choice of motif length, TCM returned more strongly conserved motifs than both Gibbs sampling and Consensus. This can be explained by the fact that TCM scores all starting positions in every sequence during each iteration, and is stochastic only in the choice of an initial profile matrix. It employs significantly more computation than either Gibbs sampling or Consensus and is able to find more strongly conserved patterns as a result. On the other hand, the Gibbs sampling algorithm depends on both a random set of initial starting positions and probabilistic choices during each iteration to select a string s_i and a new starting position within that string. Consensus is similar to TCM in that it is stochastic only in its initial choice of sequences to use as seed, but unlike TCM, where a poor initial choice can be corrected during subsequent iterations, in the case of Consensus, the effects of a poor initial choice propagate all the way through.

Although TCM produced the best results in this case, the process of scoring every starting position in each sequence was considerably more time consuming and took an order of magnitude more time than either Gibbs sampling and Consensus.

Gibbs2 and Seeded Consensus

Figure 4-20 shows the motifs discovered by the Gibbs2 algorithm with an initial working set of size 12 containing sequences chosen at random. The size of the initial working set was determined from the average contribution of each sequence to the log-odds likelihood of the best scoring motif. Figure 4-21 illustrates how the average contribution of the log-odds likelihood changed with increased values of C.

In this case, the predictive motif found once again comprised runs of premature ventricular contractions, but was more strongly conserved than the best results

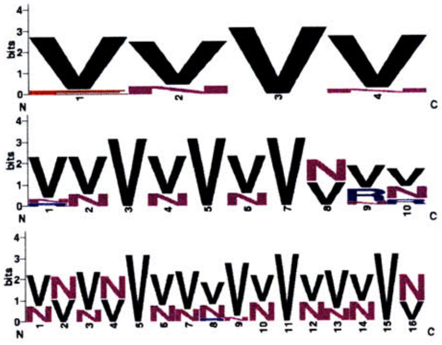


Figure 4-20: Motifs of length 4, 10 and 16 found using Gibbs2.

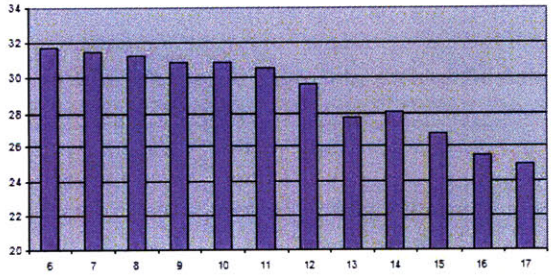


Figure 4-21: Relation of the average contribution of each sequence to the log-odds likelihood for the best scoring motif with increasing values of C.

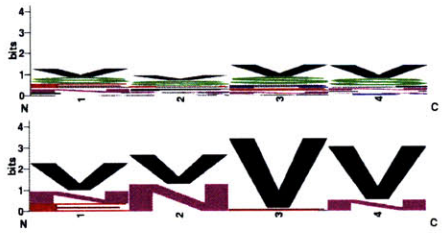


Figure 4-22: Motifs of length 4 found using Consensus (top) and Seeded Consensus (bottom).

produced earlier by TCM, Gibbs sampling and Consensus. Specifically, comparing Figures 4-17 to 4-20, the stack of symbols in Figure 7 shows the premature ventricular activity figuring more prominently at positions within the motifs.

This effect may be attributed to the ability of Gibbs2 to select a group of patients who had matching motifs comprising premature ventricular activity, unlike TCM, Gibbs sampling and Consensus, which were constrained to find a less conserved intermediate that was a best fit for data from all the different patients in the population. For this reason, Gibbs2 provided an improvement not only over the original Gibbs sampling algorithm but also the more computationally intensive TCM. The Gibbs2 algorithm has the same basic structure as the original Gibbs sampling technique, but is able to outperform TCM by addressing the issue of subsets of the population exhibiting different regulatory activity.

Figure 4-22 presents the result of using Seeded Consensus to detect motifs of length 4 relative to the original Consensus algorithm. In this case, the Gibbs2 algorithm with a working set of size 5 was used to first find an initial seed for the Consensus algorithm. As the data shows, Seeded Consensus produced more strongly conserved results than the original Consensus algorithm. This effect followed from reducing the chance that a poor initial choice of sequences would propagate and adversely affect the search for motifs.

The motif found using Seeded Consensus in Figure 4-22 is not as strongly conserved as the one discovered by Gibbs2 in Figure 4-20. This can be explained by the fact that Seeded Consensus uses Gibbs2 to discover an initial seed but otherwise

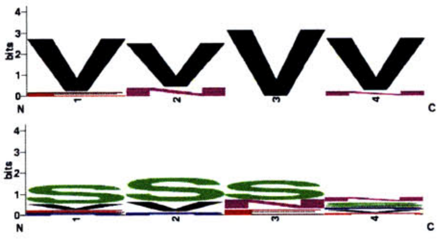


Figure 4-23: Two-stage Gibbs2 motifs of length 4. The top motif comprises a working set of size 12, while the second motif corresponds to those 11 sequences (from a total population of 23) that were not included in the original working set.

still operates on all the sequences in the data. The issue of motifs occurring only in a subset of patients does not therefore get addressed, although Seeded Consensus is still able to produce results that are comparable with TCM without the need for intensive computation.

The results of these experiments suggest that subset based techniques using Gibbs2 either to search for motifs directly, or for the purpose of providing seeds that can be fed into the Consensus algorithm, may allow for more strongly conserved motifs to be discovered than through use of TCM, Gibbs sampling and the original Consensus algorithm. Moreover, the improvement provided by the Gibbs2 algorithm proposed in our work is not associated with a significant computational overhead. In addition, the ability to partition the data into groups with homogenous motifs allows for the discovery of more than one predictive pattern, each of which may be associated with the outcome in a different group of patients. We explore this idea in more detail in the next section.

Two-Stage Gibbs2

For the motif of length 4, the sequences remaining outside the working set at the termination the of Gibbs2 algorithm were searched for a second motif common to this group. Figure 4-23 shows the results of this approach.

In this case, a second motif comprising runs of supraventricular premature or ectopic beats was found among this subgroup of the population. Notably, these patients did not show a motif similar to the ones found earlier, i.e., comprising premature ventricular beats, during any of the multiple executions of the motif discovery algorithm. This suggests that the subset of patients left outside the working set by Gibbs2 did not exhibit regulatory activity similar to the ones for whom a premature ventricular motif was discovered. Including these patients in the search for a predictive motif, as would be the case for non-subset-based techniques, would therefore lead to a less informative motif and would obscure the fact that different groups of patients show varied predictive patterns associated with an endpoint.

Motif-Event Delay

Using the motif of length 10 shown in Figure 4-20, for each sequence, the time delay between the starting location of the motif, i.e., pi, and the clinical endpoint (the occurrence of VF in the patients) was calculated for the Gibbs2 algorithm. For one of the 23 patients in the dataset, the motif occurred less than a minute prior to the event itself. In all other cases, the motif discovered preceded the actual event by at least 20 minutes or more. The median motif-event delay was 60 minutes, while the 25% and 75% quartile times were 42 and 179 minutes respectively. The maximum time separation of the motif and the event was 604 minutes.

These results suggest that the motif occurred sufficiently in advance of the endpoint to be considered merely an extension of the final event itself. Furthermore, the fact that the motif may occur at a wide range of times prior to the endpoint reinforces the need to carry out the search for predictive patterns in an automated manner, which is able to relate information across a range of positions within each sequence.

Comparison with Controls

For each patient in the SDDB population, the log-likelihood score was calculated for each starting position in the ECG label sequence. The overall score for the patient was the maximum log-likehood score found. Intuitively, this strategy assigns each patient the risk score associated with the occurrence of the discovered motif of length 10 shown in Figure 4-20 at any point during the recording, i.e., if activity similar to the motif associated with sudden death is seen anywhere, the patient is assumed to be at higher risk for the event.

Figure 4-24 shows the probability density function that can be estimated from the scores for the SDDB population. A similar strategy was adopted to score patients in two control datasets; the Physionet Normal Sinus Rhythm Database (NSRDB) and the Physionet Supraventricular Arrhythmia Database (SVDB). The decision to use SVDB data in addition to normal individuals was owing to the fact that the SVDB signals contained the same labels as the SDDB data with a higher background frequency of abnormal symbols. This ensured that a difference in scores across populations did not result from an absence of labels, but more so because activity was organized in different forms. Specifically, 1.45% of the beats in the SDDB data were premature ventricular contractions. By comparison, 5.39% of the beats in the SVDB signals and 0.002% of the NSRDB beats fell into the same category. This suggests that if the motifs seen in the SDDB population were random occurrences, then they would be expected to be seen more frequently in the SVDB dataset. With this in mind, the fact that SVDB patients had a higher percentage of premature ventricular activity but still scored lower on the discovered motifs provides further indication that the motif corresponded to activity that was not purely a random occurrence in the sudden death population. Using a maximum likelihood separator, we were able to use our motif to correctly identify 70% of the patients who suffered sudden cardiac death during 24 hours of recording while classifying none of the normal individuals, and only 8% of the patients from the supraventricular dataset as being at risk. The small number of patients in the dataset, however, does not allow for us to make statistically significant clinical statements about these findings.

4.3.3 Pattern Discovery in Positive/Negative Examples

Section 4.3.1 focused on the discovery of predictive patterns when only positive examples are available (i.e., where events occur). If negative examples (i.e., where events do not occur) are further available, this extra information can be factored into the search.

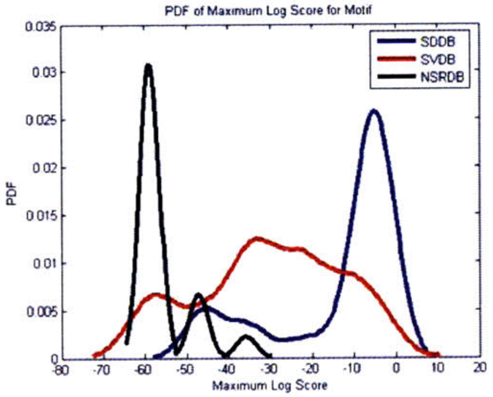


Figure 4-24: Motifs of length 4 found using Consensus (top) and Seeded Consensus (bottom).

Pattern discovery has been proposed in this context as a machine learning problem [30]:

Given two sets of sequences S^+ and S^- drawn randomly from families F^+ and F^- respectively such that $F^+ \cap F^- = \oslash$, find the pattern W of length L that has high likelihood in F^+ but not in F^-.

This formulation is sufficiently general to apply to a wide variety of applications where sequential data exists. We make the notion of a pattern more explicit by refining the goal of pattern discovery described above as follows:

Given two sets of sequences S^+ and S^- drawn randomly from families F^+ and F^- respectively such that $F^+ \cap F^- = \oslash$, find the subsequence W of length L that occurs with a Hamming distance of at most d with higher likelihood in F^+ but not in F^-.

In what follows, we propose a method to efficiently carry out the search for such approximate patterns. A variety of techniques have been proposed to address this problem statement [33, 53, 35, 38, 39, 41]. The common strategy adopted by these methods is to approach the problem of pattern discovery by finding activity that is statistically unlikely but occurs consistently in positive examples. Negative examples are primarily used for evaluation. This process means that discriminative patterns in negatively labeled sequences are not explored for classification. Other algorithms for discovery of patterns [43, 42] enumerate all exact patterns across both positive and negative examples to identify sequences that can discriminate between these two cases, but become computationally intractable when allowing subsequences to have an approximate form. As a result, sequential pattern discovery methods have traditionally been divided into two major groups: identifying approximately conserved patterns in positive examples, or finding exactly conserved patterns using both positive and negative instances.

We describe a locality sensitive (LSH) based algorithm to efficiently estimate the frequencies of all approximately conserved subsequences with a certain Hamming radius in both positive and negative examples. The search process attempts to identify patterns that allow maximum discrimination between the two groups. In this way, our method unifies the broad areas of existing work in sequential pattern detection for classification by proposing a way to discover patterns that are both approximate and derived using the additional information available in negative instances.

LSH forms a key component of our method. The use of LSH has been proposed earlier in the context of pattern discovery to identify interesting activity in positive examples [31, 32]. We supplement this work by allowing for information from negative examples to be factored into the search. In particular, we expand the use of LSH in pattern discovery from indexing to fast counting and approximate clustering. While LSH provides runtime efficiency to the search process, it imposes significant space requirements, and we describe an iterative method that uses a single LSH table in memory to address this issue. We also explore the idea of using clustering as part of pattern discovery to reduce approximate subsequences with significantly overlapping Hamming radii to a small number. This aspect of our work resembles efforts for web clustering [36]. We explore similar ideas within the context of approximate pattern discovery. This decreases the number of motifs to be evaluated while still providing a fairly exhaustive coverage of the search space. We describe a clustering method based on a 2-approximate solution of the k-center problem to achieve this goal.

The process of identifying patterns with discriminative ability makes use of concordance and rank sum testing. In many cases, the goodness of approximate patterns can be tested without using data from all training sequences. We propose a further

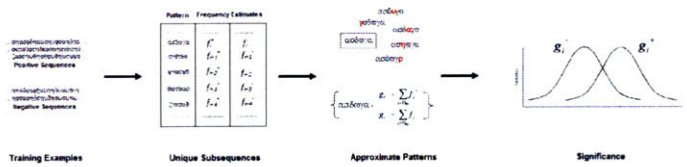


Figure 4-25: Overview of the pattern discovery process.

optimization to address these cases. The runtime and space requirements of the pattern discovery process can be reduced by using sequential statistical methods that allow the search process for patterns to terminate after using data from only as many training examples as are needed to assess significance.

Overview

The process of discovering discriminative patterns of a specified length L from positive and negative sequences is carried out in two stages: frequency estimation and pattern ranking. Figure 4-25 presents an overview of the pattern discovery algorithm.

Frequency Estimation: Given a set of positive examples $S^+ = \{S_x^+ | x = 1, ..., N^+\}$ and a set of negative examples $S^- = \{S_y^- | y = 1, ..., N^-\}$ the frequency estimation step measures the frequency of every unique subsequence W_i for $i = 1, ..., M$ in sequences belonging to S^+ and S^-. The resulting frequency for W_i in positive and negative examples is denoted as:

$$f_i^+ = \{f_{i,z}^+ | z \in S^+\} f_i^- = \{f_{i,z}^- | z \in S^-\} \tag{4.13}$$

where $f_{i,z}^+$ and $f_{i,z}^-$ are the frequencies with which W_i appears in sequences z drawn from S^+ and S^-, and f_i^+ and f_i^- are vectors measuring the frequency of W_i in all positive and negative sequences.

To allow for approximate patterns, unique subsequences are then matched to all other subsequences at a Hamming distance of at most d from W_i. Denoting this group of subsequences as D_{W_i}, the resulting frequency for the subsequence W_i and its approximate matches is defined as:

$$g_i^+ = \sum_{j \in D_{W_i}} f_j^+ g_i^- = \sum_{j \in D_{W_i}} f_j^- \tag{4.14}$$

where g_i^+ and g_i^- are vectors obtained by summing up the vectors f_i^+ and f_i^- for all subsequences within a given Hamming radius d of W_i.

In what follows, we describe an LSH-based solution that allows for efficient discovery of the subsequences D_{W_i} matching W_i. We also present a clustering approach to reduce overlapping approximate patterns for which frequencies are estimated to a smaller number with less redundancy for subsequent analysis.

Pattern Ranking: The goal of the search process is to identify approximately matching subsequences that can discriminate between positive and negative training examples. The pattern ranking stage therefore scores each candidate approximate pattern according to its discriminative ability. We use two measures to assess the goodness of patterns.

The first approach to score patterns is to use rank sum testing. This technique is a non-parametric approach for assessing whether two samples of observations come from the same distribution. Patterns are ordered based on the significance of separation (as measured by the p-value) obtained by rank sum testing. A second scoring criterion used by our work is the c-statistic, which corresponds to the area under the receiver operating characteristic (ROC) curve. Details of these techniques are provided in the remainder of this Section. We further describe how sequential methods can be used to reduce the search process to only process as many training examples as are needed to determine if a candidate pattern has high or low discriminative ability.

Locality Sensitive Hashing

Finding Approximate Matches for a Subsequence: Locality sensitive hashing [44] has been proposed as a randomized approximation algorithm to solve the nearest neighbor problem. Given a set of subsequences, the goal of LSH is to preprocess the data so that future queries searching for closest points under some l_p norm can be answered efficiently. A brief review of LSH is presented here.

Given two subsequences S_x and S_y of length L, we describe them as being similar if they have a Hamming distance of at most d. To detect similarity, we choose K indices $i_1, ..., i_K$ at random with replacement from the set $\{1, ..., L\}$. The locality sensitive hash function $LSH(S)$ is then defined as:

$$LSH(S) =< S[i_1], ..., S[i_k] > \tag{4.15}$$

where < ... > corresponds to the concatenation operator. Under this scheme, S_x and S_y are declared to be similar if:

$$LSH(S_x) = LSH(S_y) \tag{4.16}$$

The equality in 4.16 corresponds to an exact match. The approximate nature of the match is captured in the indices chosen, which may span less than the entire original subsequences S_x and S_y.

Practically, LSH is implemented by creating a hash table using the $LSH(S)$ values for all subsequences as the keys. Searching for the approximate neighbors of a query subsequence corresponds to a two-step process. The locality sensitive hash function is first applied to the query. Following this, the bucket to which the query is mapped is searched for all original subsequences with a Hamming distance of at most d.

Two sequences with a Hamming distance of d or less may not match for a random choice of K indices if one of the indices corresponds to a position in which S_x and S_y differ. The probability of such a miss is bounded by [44]:

$$Pr[LSH(S_x) \neq LSH(S_y)] \leq [1 - (1 - \frac{d}{L})^K] \tag{4.17}$$

By repeating the process of choosing K indices T times this probability can be reduced further to:

$$Pr[LSH(S_x) \neq LSH(S_y)] \leq [1 - (1 - \frac{d}{L})^K]^T \tag{4.18}$$

Effectively, 4.18 corresponds to constructing a data structure comprising T hash tables using different locality sensitive hash functions $LSH1(S), ..., LSHT(S)$. Approximate neighbors for a query are detected by searching for matches in each of these hash tables as described earlier.

The intuition behind LSH can be understood as an attempt to reduce the problem of searching through all possible sequences in the dataset for a match to the more feasible problem of searching through a much smaller set of false positives with a bounded error. The lower the desired error bound for false negatives affecting correctness (i.e., by choosing K and T), the higher the corresponding false positive rate affecting the runtime of the algorithm. The choice between these two parameters depends on the application and the underlying dataset.

Finding Approximate Matches Between All Subsequences: LSH provides an efficient mechanism to find the nearest neighbors of a given subsequence. To find the nearest neighbors for all M subsequences in the dataset, each member of the set can be passed through the entire LSH data structure comprising T hash tables for matches. Unfortunately, this process is both computationally and memory intensive. In what follows, we describe a strategy to reduce the space requirements of LSH-based search for all approximate matches between subsequences. In what follows, we address runtime issues by proposing a clustering amendment to the search process.

Different approaches have been proposed recently to reduce the space requirements of LSH. In particular, the use of multi-probe LSH [40] has been shown to substantially reduce the memory requirements for traditional LSH by searching each hash table corresponding to a random selection of K indices more thoroughly for misses. This additional work translates into fewer LSH hash tables being needed to bound the given error rate and the space of the LSH data structure decreases. In our work, the memory requirements of LSH are reduced by organizing the approximate matching process as T iterations. Each iteration makes use of a single locality sensitive hash function and maintains only a solitary hash table at any time in memory. To preserve state across iterations, the search process maintains a list of matching pairs found during each loop after removing false positives by measuring the actual distance. The subsequences D_{W_i} matching W_i are found as:

$$D_{W_i} = \bigcup_{t=1}^{T} \{W_j | LSH_t(W_j) = LSH_t(W_i)\} \tag{4.19}$$

Clustering Subsequences

The runtime of the pattern discovery process as described so far is dominated by approximate matching of all subsequences. Every subsequence is first used to create the LSH data structure, and then passed through the LSH data structure to find matches with a Hamming distance of at most d. This process is associated with considerable redundancy, as matches are sought individually for subsequences that are similar to each other. The overlap between approximate patterns increases the computational needs of the pattern discovery process and also makes it more challenging to interpret the results as good patterns may appear many times in the output.

To address this issue, we reduce patterns to a much smaller group that still collectively spans the search space. This is done by making use of a clustering method based on a 2-approximate solution to the k-center problem. The focus of this clustering is to group together the original subsequences falling into the same hash bucket during the first LSH iteration. Each of the clusters obtained at the end of this process corresponds to an approximate pattern that is retained. During subsequent iterations, while all subsequences are still used to construct the LSH tables, only the cluster centroids are passed through the LSH data structure. This reduces the runtime of the search by reducing the number of times subsequences have to be passed through the LSH tables to find true and false positives. It also reduces the memory requirements of the search by reducing the number of subsequences for which we need to maintain state about approximate matches.

The traditional k-center problem can be formally posed as follows. Given a complete graph $G = (V, E)$ with edge weights $\omega_e \geq 0$, $e \in E$ and $\omega(v, v) = 0$, $v \in V$, the k-center problem is to find a subset $Z \in V$ of size at most k such that the following quantity is minimized:

$$W(Z) = \max_{i \in V} \min_{j \in Z} \omega_{(i,j)} \tag{4.20}$$

The k-center problem is NP-hard, but a 2-approximate solution has been proposed [37] for the case where the triangular inequality holds:

$$\omega_{(i,j)} + \omega_{(j,k)} \geq \omega_{(i,k)} \tag{4.21}$$

The Hamming distance metric obeys the triangular inequality. Under this condition, the process of clustering can be decomposed into two stages. During the first LSH iteration, we identify subsequences that serve as cluster seeds using the 2-approximate solution to the k-center problem. Subsequent LSH iterations are used to grow the clusters till the probability that any subsequence within a Hamming distance at most d of the cluster centroid is missed becomes small. This approach can be considered as being identical to choosing a set of subsequences during the first LSH iteration, and finding their approximate matches by multiple LSH iterations.

More formally, during the first LSH iteration, for each bucket b_i in the hash table for $i = 1, ..., B$, we solve the k-center problem using the 2-approximate method [37] with a Hamming distance metric. The number of subsequences forming centers k_i for the i-th hash table bucket is determined alongside the specific centroid susbsequences from:

$$k_i = \min\{k | \min W(z_i(k)) \leq d\} \tag{4.22}$$

where $W(Z)$ is defined as in 4.20 and $z_{i(k)}$ denotes the subsequence centers chosen for a particular choice of k in 4.22, i.e.:

$$k_i = \min\{k | \max_{j \in b_i} \min_{z_i(k)} \omega_{(j,z_i(k))} \leq d\} \tag{4.23}$$

The final set of subsequences chosen as centroids at the end of the first LSH iteration then corresponds to:

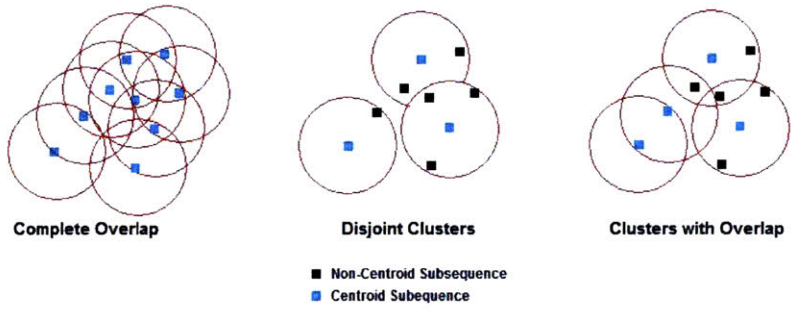


Figure 4-26: In the absence of clustering there is significant redundancy between the Hamming radii of approximate patterns. Partitioning the data into disjoint clusters can help address this issue. In our work, we reduce the original approximate patterns into a small group with some overlap to span the search space.

$$\Phi = \bigcup_{i=1}^{B} z_i(k_i) \tag{4.24}$$

The LSH iterations that follow find approximate matches to the subsequences in Φ. It is important to note that while clustering reduces a large number of overlapping approximate patterns to a much smaller group, the clusters formed during this process may still overlap. This overlap corresponds to missed approximate matches that do not hash to a single bucket during the first LSH iteration. Techniques to merge clusters can be used at the end of the first LSH iteration to reduce overlap. In our work, we tolerate small amounts of overlap between clusters analogous to the use of sliding windows to more thoroughly span the search space. Figure 4-26 illustrates the clustering process.

Pattern Ranking

Given the frequencies g_i^+ and g_i^- of an approximate pattern, corresponding to all subsequences within a Hamming distance d of the subsequence W_i, a score can be assigned to the pattern by using concordance statistics and rank sum testing.

Concordance Statistic: The concordance statistic (c-statistic) [116] measures the discriminative ability of a feature to classify binary endpoints. The c-statistic corresponds to the area under the receiver operating characteristic (ROC) curve, which describes the inherent tradeoff between sensitivity and specificity. As opposed to measuring the performance of a particular classifier, the c-statistic directly measures the goodness of a feature (in this case the frequency with which an approximate pattern occurs) by evaluating its average sensitivity over all possible specificities.

The c-statistic ranges from 0-1. A pattern that is randomly associated with the labels would have a c-statistic of 0.5. Conversely, good discriminators would correspond to either low or high c-statistic values.

Rank Sum Testing: An alternate approach to assess the goodness of patterns is to make use of rank sum testing [34, 45]. This corresponds to a non-parametric method to test whether a pattern occurs with statistically different frequencies in both positive and negative examples.

Given the frequencies g_i^+ and g_i^- of an approximate pattern in both positive and negative examples, the null and alternate hypotheses correspond to:

$$H_0 : \mu_{g_i^+} = \mu_{g_i^-} H_1 : \mu_{g_i^+} \neq \mu_{g_i^-} \tag{4.25}$$

Rank sum testing calculates the statistic U whose distribution under H_0 is known. This is done by arranging the g_i^+ and g_i^- into a single ranked series. The ranks for the observations from the g_i^+ series are added up. Denoting this value as R^+ and the number of positive examples by N^+, the statistic U is given by:

$$U = R^+ - \frac{N^+(N^+ + 1)}{2} \tag{4.26}$$

The obtained value of U is compared to the known distribution under H_0 and a probability for this observation corresponding to the null hypothesis is obtained (i.e., the p-value) under the known distribution. For large samples, U is approximately normally distributed and its standardized value can be checked in tables of the normal distribution [48, 47]. The lower the p-value obtained through rank sum testing, the more probable it is that g_i^+ and g_i^- have distributions with different means, i.e., that the approximate pattern is distributed differently in positive and negative examples.

Sequential Statistical Tests: The runtime and space requirements of the pattern discovery process can be reduced by analyzing only a subset of the training data. For example, it may be possible to recognize patterns with high discriminative ability without the need to analyze all positive and negative examples. Our pattern discovery algorithm starts out by using a small subset of the initial training data for batch analysis. This helps identify candidate approximate patterns that occur in the dataset and collectively span the search space. The remaining training examples are used for scoring purposes only. These examples are added in an online manner and during each iteration, the occurrence of outstanding candidate patterns in positive and negative examples is updated. Patterns may then be marked as being good or bad (and consequently removed from further analysis) or requiring further information to resolve uncertainty. The number of candidate patterns is therefore monotonically non-increasing with iteration number. This results in the process of analyzing additional training examples becoming faster as more data is added since fewer patterns need to be scored.

A sequential formulation for rank sum testing has been proposed [29] that adds positive and negative examples in pairs. The frequencies of an approximate pattern in positive and negative examples at the end of iteration n can be denoted as $g_i^+(j)$ and $g_i^-(j)$ where $j = 1, ..., n$. The corresponding statistic for rank sum testing is:

$$U_n = \sum_{x=1}^{n} \sum_{y=1}^{n} I(g_i^+(x) > g_i^-(y)) \tag{4.27}$$

The operator $I(.)$ is equal to one when the inequality holds and zero otherwise. Using this statistic, the decision at the end of the n-th iteration corresponds to accepting H_0 if:

$$\frac{U_n}{n} < \frac{n}{2} - \lambda \log(\frac{1-\beta}{\alpha}) \tag{4.28}$$

while H_1 is accepted if:

$$\frac{U_n}{n} > \frac{n}{2} - \lambda \log(\frac{\beta}{1-\alpha}) \tag{4.29}$$

where λ is defined as [29]:

$$\lambda = \frac{1 - \frac{\delta^2}{2} - \frac{\delta^3}{3\sqrt{3}} - \frac{\delta^4}{48}}{2\sqrt{3}\delta - \frac{\delta^2}{2}} \tag{4.30}$$

In 4.28 to 4.30, α and β correspond to desired false positive and false negative rates for the sequential rank sum testing process, while δ is a user-specified parameter that reflects the preferences of the user regarding how fine or coarse a difference the distributions are allowed to have under the alternate hypothesis. If both inequalities are not met, the process of adding data continues until there are no more training examples to add. In this case, all outstanding candidate patterns are rejected.

This formulation of sequential rank sum testing adds data in pairs of positive and negative examples. In cases where there is a skew in the training examples (without loss of generalization we assume a much larger number of negative examples than positive ones), we use a different formulation of sequential testing [46]. Denoting the mean frequency in positive training samples as:

$$\mu_i^+ = \sum_{x=1}^{N^+} g_i^+(x) \tag{4.31}$$

The alternate hypothesis can be redefined as the case where $h_\imath = g_\imath^- - \mu_i^+$ is asymmetrically distributed about the origin. This can be identified using the statistic:

$$U_n = \sum_{x=1}^{n} \frac{1}{x+1} \mathrm{sgn}(h_i(x)) R_{xx} \tag{4.32}$$

where R_{xy} is defined as the rank of $|h_\imath(x)|$ in the set $\{|h_\imath(1)|, ..., |h_\imath(y)|\}$ with $x = y$ and $\mathrm{sgn}(|h_\imath(x)|)$ is 1 if $h_\imath(x) = 0$ and -1 otherwise. The test procedure using the statistic continues taking observations as long as $U_n \in (-\delta, \delta)$ where δ is a user-specified parameter.

Traditional formulations of sequential significance testing remove good patterns from analysis when the test statistic is first found to lie above or below a given threshold. Patterns with potentially low discriminative ability are retained for further analysis and are discarded if they do not meet any of the admission criteria during any iteration of the search process. Since there may be a large number of such patterns with low discriminative value, we make use of a modified approach to reject poor hypotheses while retaining patterns that may have value in classification. This strategy improves efficiency, while also ensuring that good patterns are ranked using the available training data. Given the typical goal of pattern discovery to return the best patterns found during the search process, this technique naturally addresses the problem statement.

Given the test statistics in 4.27 and 4.32, we remove all patterns that have a test statistic:

$$U_n < \lambda U_{\max} \tag{4.33}$$

where $U_{\max}$ is the maximum test statistic for any pattern and λ is a user-specific fraction (e.g., 0.2).

Multiple Hypothesis Correction: While assessing a large number of approximate patterns, M, the statistical significance required for goodness must be adjusted for Type I (i.e., false positive) errors. If we declare a pattern to be significant for some probability of the null hypothesis less than θ, then the overall false positive rate for the experiment assuming independence of patterns is given by:

$$FP = 1 - (1 - \theta)^M \tag{4.34}$$

If we do not assume that the patterns are independent, the false positive rate can be bounded by:

$$FP \leq \theta M \tag{4.35}$$

To account for this condition, a more restrictive level of significance must be set consistent with the number of unique patterns being evaluated (in our case, the clusters obtained earlier). If c clusters are assessed for goodness, the Bonferroni correction [28] suggests that the level of significance be set at:

$$\theta' = \frac{\theta}{c} \tag{4.36}$$

This addresses the issue of increasing false positives caused by the evaluation of a large number of clusters by correspondingly lowering the p-value required to accept a pattern.

4.3.4 Evaluation of Pattern Discovery in Positive/Negative Examples

We evaluated our method on 24-hour electrocardiographic (ECG) signals from the DISPERSE2 TIMI33 and MERLIN TIMI36 trials (Chapter 3.3). These data were collected from patients admitted with non-ST-elevation acute coronary syndromes (NSTEACS). We applied our pattern discover algorithm to discover specific sequences of beat-to-beat morphology changes that had predictive value for future cardiovascular death.

Given a 24-hour ECG signal, we first converted the data recorded during the course of hospitalization into a morphology differences (MD) time series using the methods described in Section 3.2.4. The morphology change time series was then converted into a sequence by using symbolic aggregate approximation (SAX) [21] with an alphabet size of 10. In this manner, multiple ECG signals were transformed into sequences that could be analyzed by our method. On average, each sequence corresponding to 24 hours of ECG was almost 100,000 symbols long.

We used the DISPERSE2 TIMI33 dataset as training data. The 15 patients who died over the 90 day period following NSTEACS were treated as positive examples, while the remaining 750 patients comprised negative examples. we applied our pattern discovery algorithm to learn sequences of morphology changes in the ECG signal that were predictive of death. We searched for patterns of length 8 with a Hamming distance of at most 2. Parameters were chosen using the inequality in Equation 4.18 so that the LSH probability of false negatives was less than 0.01. We selected patterns that showed a c-statistic greater than 0.7 and a rank sum test p-value of 0.05 corrected for multiple hypotheses using the Bonferroni correction. These were then evaluated on a test set of 250 patients from the MERLIN TIMI36 trial with 10 deaths.

We also studied the runtime performance of our algorithm on this dataset with and without the use of sequential statistics to find significant patterns. Given the large number of negative examples in the training data, we employed the sequential formulation in Equations 4.31 and 4.32 with $\lambda = 0.2$. We denote our original algorithm using sequential statistics as *LSHCS* while the variation that avoids sequential statistics is denoted by *NoSeqStats*.

Our pattern discovery method returned 2 approximate patterns that were assessed to have discriminative value in the training set (i.e., a c-statistic of more than 0.7 and a p-value of less than 0.05 after accounting for the Bonferroni correction). Representing the symbols obtained using SAX by the letters A-J, where A corresponds to the symbol class for the least beat-to-beat change in morphology and J denotes the symbol for the greater change, the centroids for the approximate pattern can be written as:

$$ABCCDFGJFFJJJJCC \tag{4.37}$$

The first of these patterns is equivalent to increasing time-aligned energy changes between successive beats. This may suggest increased instability in the conduction system of the heart. The second pattern corresponds to a run of instability followed by a return to baseline. This pattern can be interpreted as a potential arrhythmia. The results of testing both patterns on previously unseen data from 250 patients (with 10 deaths over a 90 day follow-up period) are shown in Table 4.4. Both patterns found by our approximate pattern discovery algorithm showed statistical significance in predicting death according to both the c-statistic and rank sum criteria in the test population.

A comparison of the running times for the *LSHCS* and *NoSeqStats* algorithms is

Table 4.4: Statistical significance of approximate patterns found on a training set of 765 post-NSTEACS patients (15 deaths over a 90 day follow-up period) when evaluated on a test population of 250 patients (10 deaths).

Pattern (Centroid)	Rank Sum P-Value	C-statistic
ABCCDFGJ	0.025	0.71
FFJJJJCC	0.004	0.70

Table 4.5: Time taken by the *LSHCS* and *NoSeqStats* pattern discovery algorithms on the cardiovascular training dataset.

Algorithm	Time
LSHCS	5:08:24
NoSeqStats	9:43:09

presented in Table 4.5. While the outputs produced by both algorithms were identical, the use of sequential statistics helped our *LSHCS* method decrease the runtime of the search process to almost half. We also note that the *NoSeqStats* variation used considerably more memory than the *LSHCS* approach. This effect was due to *LSHCS* purging state for patterns that did not obey the inequality in Equation 4.32. In the absence of sequential statistics, *NoSeqStats* had to retain ranking information for all patterns till the training dataset was completely analyzed.

4.3.5 Symbolic Mismatch

In Sections 4.3.1 and 4.3.3, we described how risk stratification could be carried out within a symbolic framework by searching for high risk symbolic patterns. Patients whose ECG signals match these patterns can be recognized as being at an elevated risk of adverse outcomes. From this perspective, the work in Sections 4.3.1 and 4.3.3 corresponds to a feature-based approach for risk stratification (where the feature of interest is whether the high risk symbolic pattern occurred or not). We now focus on an alternative approach. In contrast to recognizing patients at high risk depending on a specific feature, we recognize high risk patients as individuals who represent outliers in a population, i.e., we develop a comparative risk stratification approach rather than a feature-based one.

Evidence suggests that high risk patients constitute a small minority. For example, cardiac mortality over a 90 day period following ACS was reported to be 1.79% for the SYMPHONY trial involving 14970 patients [27] and 1.71% for the DISPERSE2 trial with 990 patients [88]. The rate of myocardial infarction (MI) over the same period for the two trials was 5.11% for the SYMPHONY trial and 3.54% for the DISPERSE2 trial.

While many different feature-based approaches have been proposed to identify high risk patients, we focus instead on finding cases that are atypical in morphology and dynamics. We propose a new metric, called the *symbolic mismatch* (SM), that quantifies the extent to which the long-term ECG recordings from two patients differ. The pairwise differences are used to partition patients into groups with similar ECG characteristics and potentially common risk profiles.

Our hypothesis is that those patients whose long-term electrocardiograms did not match the dominant group in the population, are at increased risk of adverse cardiovascular events. These cases have a high symbolic mismatch relative to the majority of the patients in the population, and form one or more subgroups that are suspected to be at an increased risk of adverse events in the future.

Our approach is orthogonal to the use of specialized high risk features along two important dimensions. Firstly, it does not require the presence of significant prior knowledge. We only assume that ECG signals from patients who are at high risk differ from those of the rest of the population. There are no specific assumptions about the nature of these differences. Secondly, the ability to partition patients into groups with similar ECG characteristics and potentially common risk profiles allows for a more fine-grained understanding of a how a patients future health may evolve over time. Matching patients to past cases with similar ECG signals could lead to more accurate assignments of risk scores for particular events such as death and MI.

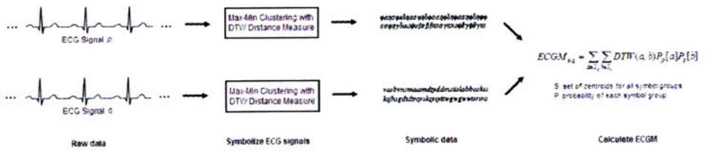

ECG Signal p
ECG Signal q
Max-Min Clustering with DTW Distance Measure
Max-Min Clustering with DTW Distance Measure
$ECGM_{p,q} = \sum_{a \in S_p} \sum_{b \in S_q} DTW(a,b) P_p[a] P_q[b]$
S: set of centroids for all symbol groups
P: probability of each symbol group
Raw data
Symbolize ECG signals
Symbolic data
Calculate ECGM

Figure 4-27: Calculating symbolic mismatch (SM) between two patients. ECG signals are first symbolized using a Max-Min iterative clustering approach that employs a dynamic time-warping (DTW) distance measure to compare beat morphology. The resulting symbol centroids and probability distribution over all symbol classes are used to obtain a final SM value measuring the long-term electrocardiographic dissimilarity between the patients.

Quantifying Differences in Symbol Distributions

The symbolic mismatch (SM) between two patients, p and q, is calculated using the process shown in Figure 4-27.

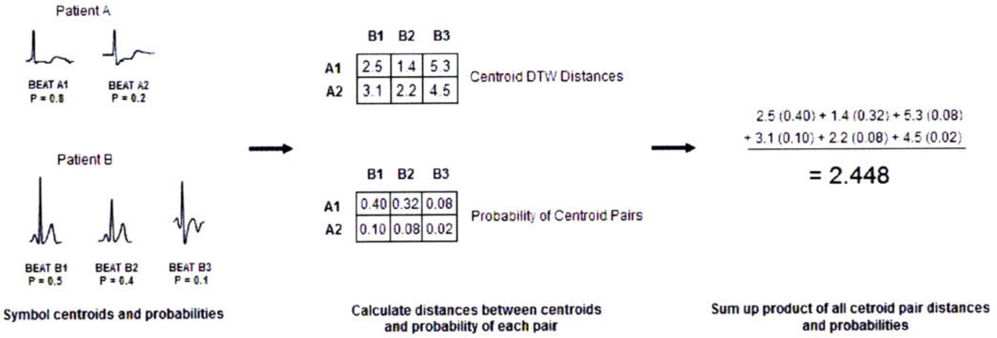
Denoting the set of symbols for patient p as S_p and the set of probabilities with which these symbols occur in the electrocardiogram as P_p (for patient q an analogous representation is adopted), we calculate the SM between these patients as:

$$SM_{p,q} = \sum_{a \in S_p} \sum_{b \in S_q} C(a,b) P_p[a] P_q[b] \tag{4.38}$$

$C(a, b)$ corresponds to the dynamic time-warping cost of aligning the centroids of symbol classes a and b.

Intuitively, the symbolic mismatch between patients p and q corresponds to an estimate of the expected dynamic time-warping cost of aligning any two randomly chosen beats from these patients. The SM calculation above achieves this by weighting the cost between every pair of symbols between the patients by the probabilities with which these symbols occur. An example of the SM calculation is presented in Figure 4-28.

An important feature of SM is that it is explicitly designed to avoid the need to set up a correspondence between the symbols of patients p and q for comparative purposes. In contrast to cluster matching techniques [25, 26] that compare data for



	B1	B2	B3
A1	2.5	1.4	5.3
A2	3.1	2.2	4.5

	B1	B2	B3
A1	0.40	0.32	0.08
A2	0.10	0.08	0.02

Figure 4-28: A hypothetical example of the SM calculation.

two patients by first making an assignment from symbols in one patient to the other, SM does not require any cross-patient registration of symbols and performs weighted comparisons between all symbols for p and q.

Hierarchical Clustering of Patients Using SM

For every pair of patients in a population, the symbolic mismatch between them is computed using the techniques described in Section 4.3.5. The resulting divergence matrix, D, relating the pairwise symbolic mismatches between all the patients is used to partition the population into groups with similar cardiac characteristics. This process is carried out by means of hierarchical clustering [24].

Hierarchical clustering starts out by assigning each patient to a separate cluster. It then proceeds to combine two clusters at every iteration, choosing clusters that obey some concept of being the "closest" pair. We use a definition of closest that corresponds to merging two clusters A and B for which the mean symbolic mismatch between the elements of the clusters is minimized, i.e., we choose clusters A and B such that they minimize the merge distance, f, which is given by:

$$f = \frac{1}{|A|.|B|} \sum_{x \in A} \sum_{y in B} SM_{x,y} \tag{4.39}$$

where $|A|$ and $|B|$ correspond to the number of elements in each cluster.

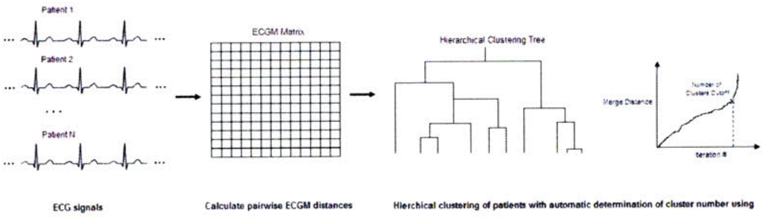

Patient 1
Patient 2
Patient N
ECG signals
ECGM Matrix
Calculate pairwise ECGM distances
Hierarchical Clustering Tree
Merge Distance
Number of Clusters Cutoff
Iteration #
Hierchical clustering of patients with automatic determination of cluster number using

Figure 4-29: Stages in the patient clustering process to determine high risk minority groups that are population outliers.

Intuitively, this approach picks two clusters to merge that are closest in the sense that the average distance between elements in the two clusters is minimized. This definition of closest is similar to the unweighted pair group method with arithmetic mean (UPGMA) or average linkage criterion [23].

Broadly speaking, there are two approaches to decide when to terminate the iterative clustering process. The simplest approach is to terminate at the iteration when the clustering process has produced a pre-determined number of clusters. However, in this case we have no prior assumptions about the appropriate number of clusters. We therefore use a more complex approach in which the number of clusters is determined by the dataset.

The merge distance defined in 4.39 is monotonically nondecreasing with iteration number. Small increases in the merge distance suggest that the clustering process is merging clusters that are close. Conversely, large merge distances correspond to clusters being merged that are dissimilar. We therefore use the merge distance to indicate when the clustering process is beginning to show diminishing returns, i.e., merging clusters that are increasingly far apart. Continuing beyond this point may lead to the new clusters created containing heterogenous elements. We therefore terminate the clustering process when the merge distance for the next three iterations would show a quadratic concave up increase.

The process of hierarchically clustering patients using SM is shown in Figure 4-29.

4.3.6 Evaluation of Symbolic Mismatch

We tested symbolic mismatch on patients in the DISPERSE2 TIMI33 trial (Sections 3.2.5 and 4.3.4). To evaluate the ability of symbolic mismatch to identify patients at increased risk of future cardiovascular events, we first separated the patients into a dominant normal sub-population (i.e., the low risk SM group) and a group of abnormal patients (i.e., the high risk SM group). This was done by terminating hierarchical clustering automatically and labeling all patients outside the largest cluster as being abnormal and potentially high risk. In the subsequent discussion, we denote this new risk variable as the ECG Non-Dominance (ECGND). Patients placed in the non-dominant group by SM clustering were assigned an ECGND value of 1, while those in the dominant group had a value of 0.

Kaplan-Meier survival analysis was used to study the event rates for death and MI. Hazard ratios (HR) and 95% confidence interval (CI) were estimated by using a Cox proportional hazards regression model to study event rates in patients within the dominant and non-dominant groups. The HR for the dominant and non-dominant SM groups was compared to other clinical risk variables; age, gender, smoking history, hypertension, diabetes mellitus, hyperlipidemia, coronary heart disease (CHD), prior MI, prior angina and ST depression on holter. The risk variables were also examined using multivariate analysis. The outcomes of death and MI were studied both separately, as well as after being combined to create a combined endpoint of death or MI (death/MI).

The results of univariate analysis for death, MI and the combined outcome are shown in Tables 4.6 to 4.8 for all risk variables including ECGND. The corresponding Kaplan-Meier curves are presented in Figure 4-30.

Of the clinical risk variables examined, age showed a consistent, though small, association with both adverse endpoints on univariate analysis over the 90 day follow-up period. The presence of diabetes was also found to be associated with cardiac mortality.

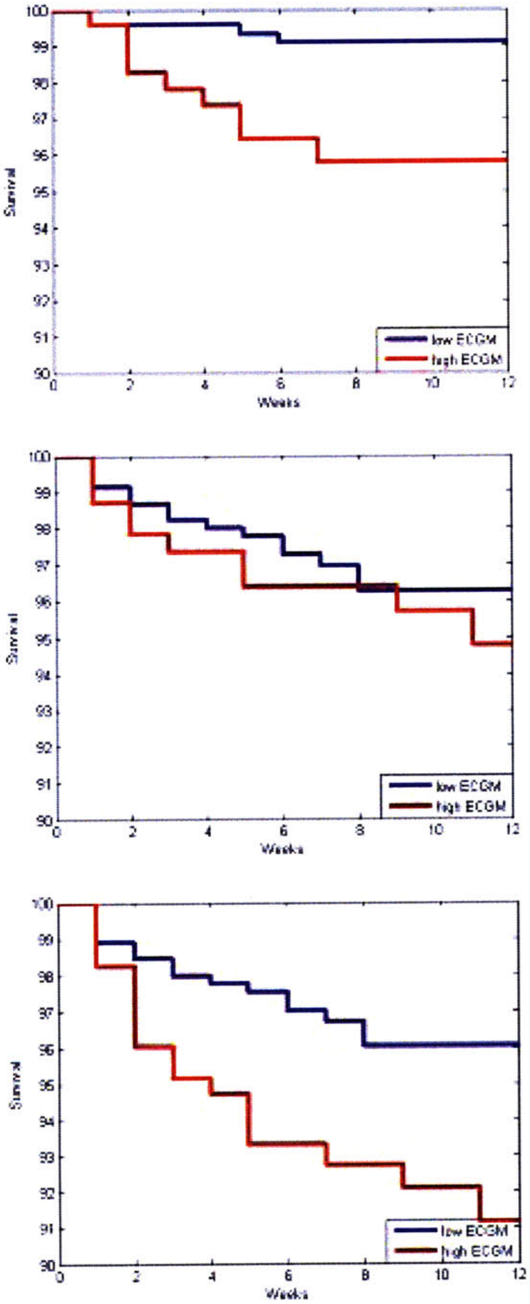
In the case of ECGND, patients who were electrocardiographically mismatched

Table 4.6: Association of risk variables with death in univariate and multivariate analyses (n=686).

Variable	Univariate Hazard Ratio	P Value	Multivariate Hazard Ratio	P Value
Age	1.10	<0.01	1.07	0.056
Gender	2.53	0.086	1.81	0.353
Smoker	0.59	0.321	2.32	0.227
Hypertension	6.03	0.083	3.39	0.253
Diabetes	3.36	0.024	1.75	0.331
Hyperlipidemia	0.57	0.288	0.61	0.411
CHD	0.13	0.051	0.23	0.164
Prior MI	2.25	0.134	2.01	0.235
Prior angina	2.61	0.141	1.83	0.393
ST depression	2.51	0.120	1.26	0.707
ECGND	4.71	<0.01	3.62	0.038

Table 4.7: Association of risk variables with MI in univariate and multivariate analyses (n=686).

Variable	Univariate Hazard Ratio	P Value	Multivariate Hazard Ratio	P Value
Age	1.04	0.034	1.05	0.011
Gender	0.50	0.128	0.38	0.050
Smoker	1.22	0.605	1.15	0.745
Hypertension	1.71	0.246	1.86	0.208
Diabetes	1.36	0.463	1.19	0.693
Hyperlipidemia	0.76	0.466	0.81	0.610
CHD	1.15	0.725	1.49	0.333
Prior MI	1.21	0.644	1.20	0.689
Prior angina	0.70	0.351	0.62	0.245
ST depression	0.73	0.411	0.61	0.238
ECGND	1.69	0.167	1.66	0.193



(a) Death (HR 4.71, p<0.01)

(b) MI (HR 1.69, p=0.167)

(c) Death/MI (HR 2.58, p<0.01)

Figure 4-30: Kaplan-Meier survival curves for (a) death, (b) MI and (c) death/MI comparing the high SM risk (n=229) and low SM (n=457) groups.

Table 4.8: Association of risk variables with death/MI in univariate and multivariate analyses (n=686).

Variable	Univariate Hazard Ratio	P Value	Multivariate Hazard Ratio	P Value
Age	1.05	<0.01	1.05	<0.01
Gender	0.77	0.468	0.63	0.242
Smoker	1.15	0.670	1.52	0.285
Hypertension	1.99	0.103	1.81	0.185
Diabetes	1.84	0.077	1.41	0.350
Hyperlipidemia	0.74	0.362	0.76	0.429
CHD	0.85	0.635	1.14	0.720
Prior MI	1.46	0.281	1.33	0.451
Prior angina	1.03	0.932	0.87	0.707
ST depression	1.04	0.910	0.74	0.387
ECGND	2.58	<0.01	2.43	<0.01

with the dominant group of the population showed an increased risk of adverse cardiovascular events. Patients outside the dominant cluster had a much higher rate of death during follow-up than patients in the dominant cluster (4.37% vs. 0.88%; p<0.01). A similar trend was seen for MI (5.68% vs. 3.28%) although in this case the relationship was not statistically significant (p=0.167). For the combined death/MI endpoint, i.e., the occurrence of either of these adverse outcomes, the cumulative incidence in the high risk group was 9.17% as opposed to 3.50% in the low risk group (p<0.01).

The results of multivariate analysis for death, MI and the combined outcome are also shown in Tables 4.6 to 4.8 for all risk variables including ECGND.

On multivariate analysis, ECGND was the only risk variable evaluated that showed an independent association with death over a 90 day period following non-ST segment elevation acute coronary syndrome (NSTEACS). There was no statistically significant association between ECGND and MI. For the combined death/MI endpoint both age and ECGND showed an independent association with the 90 day outcome, with a higher hazard ratio being observed in the case of ECGND.

Hierarchical patient clustering produced 46 clusters in all, i.e., one dominant clus-

Table 4.9: Percentage of patients with events in five largest clusters in the high SM risk group relative to low SM risk group.

Cluster	# of Patients	%	% MI	%
		Death		Death/MI
A	53	3.77	1.89	5.66
B	48	2.08	8.33	10.42
C	22	18.18	4.55	22.73
D	20	10.00	5.00	10.00
E	12	0.00	16.67	16.67
Low SM	457	0.88	3.28 3.50	

ter that constituted the low risk SM group of patients and 45 clusters that collectively formed the high risk SM group. Of the high risk SM clusters, 31 had only a single element, potentially corresponding to isolated singletons resulting from noisy electrocardiograms. Conversely, 5 of the high risk SM clusters had 10 or more patients. These 5 clusters are labeled A-E.

Table 4.9 presents the risk of events for the non-dominant clusters comprising 10 or more patients. The data suggest that patients in different clusters may have distinct risk profiles. Consider, for example, patients in the C cluster. The risk of death for these patients is 18.18% relative to a risk of 0.88% in the dominant cluster population (HR 23.20, $p<0.01$). The overall risk of death/MI is also correspondingly elevated (22.73% vs. 3.50%; HR 7.54, $p<0.01$). Cluster C had a higher rate of death not only than the dominant cluster, but also relative to other non-dominant clusters (HR 9.05, $p<0.01$) suggesting a worse prognosis for patients in that group.

Similarly, in the E cluster, there are no deaths but the risk of MI is 16.67% as opposed to 3.28% in the dominant cluster (HR 5.35, $p=0.026$). However, in this case, there is no statistically significant increase in the rate of MI relative to the rest of the non-dominant clusters (HR 2.99, $p=0.164$).

The percentage of events in the combined population comprised only by the non-dominant clusters with 10 or more members (i.e., A-E) is shown in Table 4.10. For each endpoint, there is a higher percentage of events in the cumulative population

Table 4.10: Percentage of patients with events in aggregate of five largest clusters in high SM risk group compared to low SM risk group.

Cluster	# of Patients	% Death	% MI	% Death/MI
A-E	155	5.81	5.81	10.97
Low SM	457	0.88	3.28	3.50

comprising only clusters with 10 or more patients than when the entire non-dominant cluster population (including clusters less than 10 patients) is analyzed. These data suggest that improved noise removal techniques, or disregarding small electrocardiographically mismatched clusters, could allow for a further focus on high risk cases.

4.4 Other Applications of Symbolic Analysis: Fetal Risk Stratification

In addition to our work on risk stratification following NSTEACS, we also applied symbolic analysis to fetal ECG data.

Inflammatory conditions such as intrauterine infection (chorioamnionitis) during pregnancy are associated with an increased risk of sepsis, cerebral palsy, and death in newborns [10, 11, 12]. Early detection of inflammation may allow for interventions that reduce the risk of adverse newborn outcome. The standard approaches to diagnosing intrauterine infection are based on measuring the mothers body temperature periodically during labor or measuring fetal HRV. These approaches only detect infections once the inflammatory process is sufficiently advanced to elevate maternal core temperature or to cause a fetal systemic response.

We propose morphologic entropy in the fetal electrocardiogram signal as a new risk metric for the early detection of inflammation and neuronal injury during pregnancy. This metric is computed using a two-step process. We first represent the original ECG signal $x[n]$ as a symbolic sequence comprising labels derived from an unsupervised algorithm to partition ECG beats into distinct classes with characteristic morphology. We then measure the entropy in this symbolic representation to derive the morphologic entropy as:

$$H(x) = -\sum_{c_i \in S} f(c_i)\log(f(c_i)) \tag{4.40}$$

where $f(c_i)$ is the frequency of c_i in the symbolic representation of the fetal ECG signal.

We evaluated morphologic entropy in a preliminary study on fetal ECG signals from five subjects. These signals were sampled at 1000 Hz with 32 bit quantization and recorded using a fetal scalp electrode placed for a clinical reason following amniotic rupture. The recording of patient data was carried out at the Brigham and Womens Hospital, Boston, MA USA, with informed consent obtained from mothers considered at high risk for delivering a baby with fetal injury. Each recording was between 57-200 minutes long with a mean recording duration of 144 minutes. We assessed the quality of each fetal ECG signal using the Physionet Signal Quality Index (SQI) package [82] and by measuring the standard deviation (SD) of the normalized R-wave amplitude. All five recordings were found to be sufficiently high quality (i.e., SQI ¿ 90% and SD ¡ 0.2887) for further analysis.

For each patient, IL-6, IL-8 and NSE were also measured from cord serum using fully-automated random and enzyme-linked immunosorbent assays. The sensitivity and coefficient of variation (CV) for the assays were 1 pg/ml and ¡ 10% for IL-6, 10 pm/ml and ¡ 10% for IL-8 and 1 ug/l and ¡ 5% for NSE. Abnormal ranges for the biomarkers were chosen from the literature to be ¿ 11 for IL-6, ¿ 90 for IL-8 and ¿ 12.5 for NSE [7, 8, 9].

Figures 4-31 - 4-33 show the association between morphologic entropy and IL-6, IL-8 and NSE. In each case, we observed a strong linear relation between morphologic entropy and marker levels in cord serum (p ¡ 0.05). As the measured IL-6, IL-8 and NSE levels increased, there was an associated increase in the entropy of the fetal ECG morphology.

In addition to the markers of inflammation and neuronal injury, periodic mater-

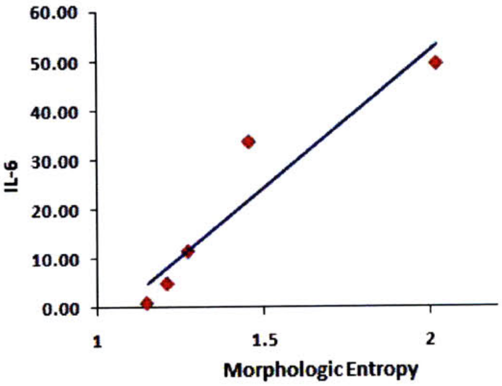


Figure 4-31: Relation between morphologic entropy and IL-6 levels in cord blood (Y = -59.13 + 55.67X; p = 0.019; standard error for coefficients = 17.38 and 11.93; RMSE = 7.68)

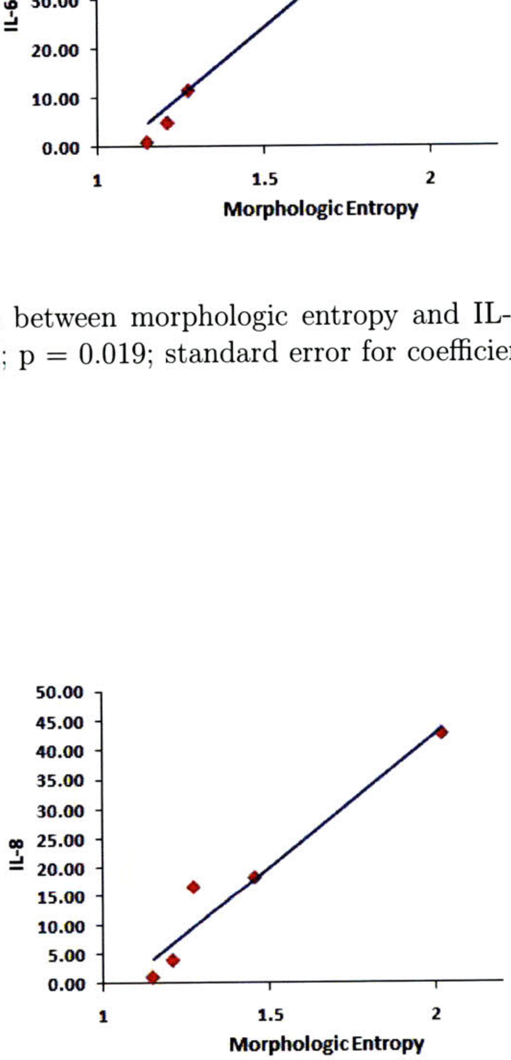


Figure 4-32: Relation between morphologic entropy and IL-8 levels in cord blood (Y = -48.89 + 45.82X; p = 0.009; standard error for coefficients = 11.01 and 7.56; RMSE = 4.75)

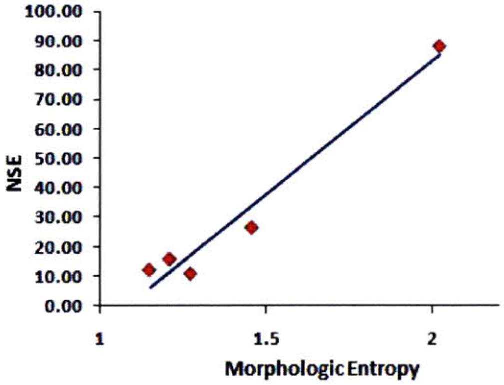


Figure 4-33: Relation between morphologic entropy and NSE levels in cord blood (Y = -97.73 + 90.38X; p = 0.005; standard error for coefficients = 17.67 and 12.14; RMSE = 7.34)

nal temperature measurements were also recorded for all five subjects. None of the mothers developed a fever during labor, despite the increased IL-6, IL-8 and NSE levels in some of the cases. Furthermore, in the period of hospitalization post-labor, fever was observed in only one of the five mothers. The cord levels of the different markers for this case were IL-6 = 4.98 pg/ml, IL-8 = 3.81 pg/ml and NSE = 15.85 ug/l, i.e., the mother did not represent one of the cases with the highest inflammatory or brain injury markers in the cord labs. This data suggests that while morphologic entropy of the fetal ECG is strongly associated with IL-6, IL-8 and NSE levels in the cord blood, the absence of fever in the mother is a poor predictor of the lack of inflammation or neuronal injury.

We also evaluated different metrics based on heart rate variability for association with IL-6, IL-8 and NSE. We measured the SDNN, SDANN, ASDNN, rMSSD, HRVI, pNN50 and LF/HF metrics for each patient. Tables 4.11 and 4.12 present the HRV metrics computed for each subject and the measured levels of the markers in cord blood. None of the HRV metrics showed a statistically significant linear relation (i.e., p ¡ 0.05) with IL-6, IL-8 or NSE. These data suggest that in this study population, HRV metrics were a poor indicator of inflammation or brain injury to the fetus.

Our findings suggest that morphologic entropy may have value in early detection of inflammatory processes such as intrauterine infections and fetal brain injury. Our

Table 4.11: HRV metrics for subjects. Mean heart rate (Mean HR) and the standard deviation of the heart rate (STD HR) are also provided for each subject.

ID	SDNN	SDANN	ASDNN	HRVI	pNN50	RMSSD	LF/HF	MEAN HR	STD HR
1	54	30	39	16	0.25	26	1.64	123	13.9
2	65	32	49	4	0.08	21	2.24	101	16.5
3	50	28	42	9	0.13	22	2.36	114	12.8
4	23	14	18	6	0.02	9	2.79	104	5.7
5	40	19	32	7	0.16	20	2.98	107	10.9

Table 4.12: Cord blood markers for subjects.

ID	IL-6	IL-8	NSE
1	34	18	27
2	12	17	11
3	49	43	88
4	5	4	16
5	1	1	12

results also suggest that morphologic entropy may be a better detector of high-risk conditions than the present strategy of measuring maternal fever or using heart rate variability-based metrics. While the results of our work are promising, we note that our dataset presently represents a small population. Although this did not prevent us from obtaining statistically significant results, because of the strong association between morphologic entropy and the different markers measured in cord blood, these findings should be considered a preliminary study. We hope to supplement these initial data with testing on a larger population to build a stronger case for the use of morphologic entropy in clinical practice. We also note that in our study, we used ECG signals recorded with a fetal scalp electrode placed during labor. While the scalp electrode used for testing is not strictly noninvasive, as it requires the insertion of a wire electrode into the scalp of the fetus, we believe that the increased entropy in fetal ECG is independent of the specific ECG acquisition technology, and speculate that it can be measured noninvasively using fetal ECG derived from maternal abdominal leads. However, further experimentation is needed to test this hypothesis.

4.5 Summary

In this chapter, we proposed the notion of computational physiology, i.e., the analysis of large amounts of physiological data as symbolic sequences. We reviewd the concept of symbolization as it appears in different disciplines, and described how we can extend this idea to physiological signals. We explained, in particular, why symbolic analysis is an appropriate and powerful paradigm for physiological data, and what domain specific challenges need to be addressed for its use in this setting. We also proposed algorithms to symbolize many broad classes of physiological signals, and to analyze symbolic representations of physiological data with the goal of risk stratification. We presented the idea of finding high risk symbolic patterns that are conserved or absent in patients experiencing adverse events, and the idea of finding patients at risk of different adverse outcomes through a comparative approach that groups together patients with similar symbolic sequences (and potentially similar risk profiles). We evaluated both these approaches, and also described how similar ideas can be applied to fetal ECG signals to risk stratify for cerebral palsy.

The work discussed in Chapters 3 and 4 focused on the analysis of large amounts of physiological data with the goal of risk stratification. We now turn out attention towards visualizing the activity discovered by our methods, so that interesting findings can be communicated in a compact form to clinicians.

Chapter 5

Visualization of Long-Term Data

In Chapters 3 and 4, we proposed techniques for the automated analysis of large amounts of continuous data. As a complement to this work, we have also developed visualization tools to help clinicians and researchers look at information in long-term signals.

Our visualization tools are based on the symbolic framework described in Chapter 4. We make use of the data reduction provided by symbolization to compactly display physiological signals as symbolic sequences rather than waveforms. We further supplement this with the ability to drill down and visualize prototypical waveforms associated with each symbol as well as individual raw data points.

In what follows, Section 5.1 presents the design concepts and interfaces for our visualization tools to compactly display information in long-term signals. We illustrate how visualizing long-term data as symbolic sequences makes it easier to identifying interesting activity. Section 5.2 then focuses on the creation of prototypes for the visualization of each symbol class, by characterizing persistent physiological activity while removing variations due to noise and time-skew.

5.1 Visualizing Data as Symbolic Sequences

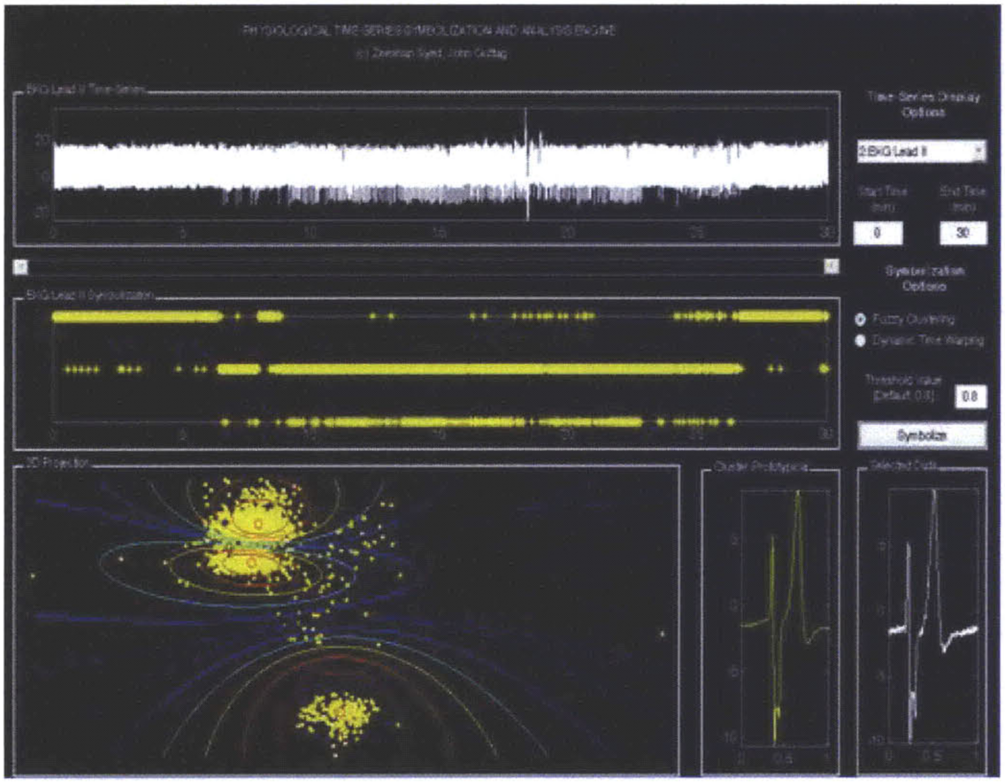
We believe that looking at data as a sequence of symbols, where each symbol has an interpretable form, helps facilitate the visualization of data. In particular, it provides many advantages over the alternative of visualzing raw samples. First, symbolization results in a large decrease in the number of data points that need to be visualized (i.e., there are fewer data points along the x-axis as many samples are compacted into a single symbol). Second, symbolization also makes it easier to see when changes occur (i.e., there are fewer discrete values for symbols than raw samples, so data is easier to visualize along the y-axis). Finally, symbolization also makes the data more readily interpretable (i.e., it may be more natural to reason in terms of changes in different classes of physiological activity than changes in terms of raw samples). Our philosophy is therefore to visualize long-term data as symbols, while retaining information (in the form of prototypes and even the raw data) that can be available to users interested in looking at waveforms.

Figure 5-1 shows an interface for our visualization tool. The raw ECG signal is shown in the top panel with a symbolic representation of this data just below. While it is hard to appreciate the different kinds of activity taking place by looking at the raw ECG signal, the symbolic display makes it easy to recognize that three different classes of heart beats occurred and also the temporal distribution of this activity.

The lower left panel of the tool shows a 2-dimensional down-projection of all the ECG beats using PCA [4]. This is intended to allow users to appreciate the intra- and inter-cluster distance between beats. Individual heart beats, and the prototypical representations of the symbol classes to which these beats were assigned, can be seen in the panels on the bottom right.

In this way, the display in Figure 5-1 provides a quick and compact way to visualize the different kinds of heart beats that were observed during the course of continuous monitoring, the time-distribution of these different classes of activity, how different these groups of beats were, and also what the prototypical representation of each beat group is.

We also allow users to choose between different clustering options for symbolization, and to configure the granularity of the clustering (top right panel). This provides users with flexibility on the process through which symbols are created, and also how many symbols to create.


PHYSIOLOGICAL TIME-SERIES SYMBOLIZATION AND ANALYSIS ENGINE
(c) Zeeshan Syed, John Guttag
EKG Lead II Time-Series
20
0
-20
0
5
10
15
20
25
30
Time-Series Display Options
EKG Lead II Symbolization
Symbolization Options
3D Projection
Cluster Prototypes
Selected Data
Symbolize

2:EKG Lead II

Start Time (min) 0

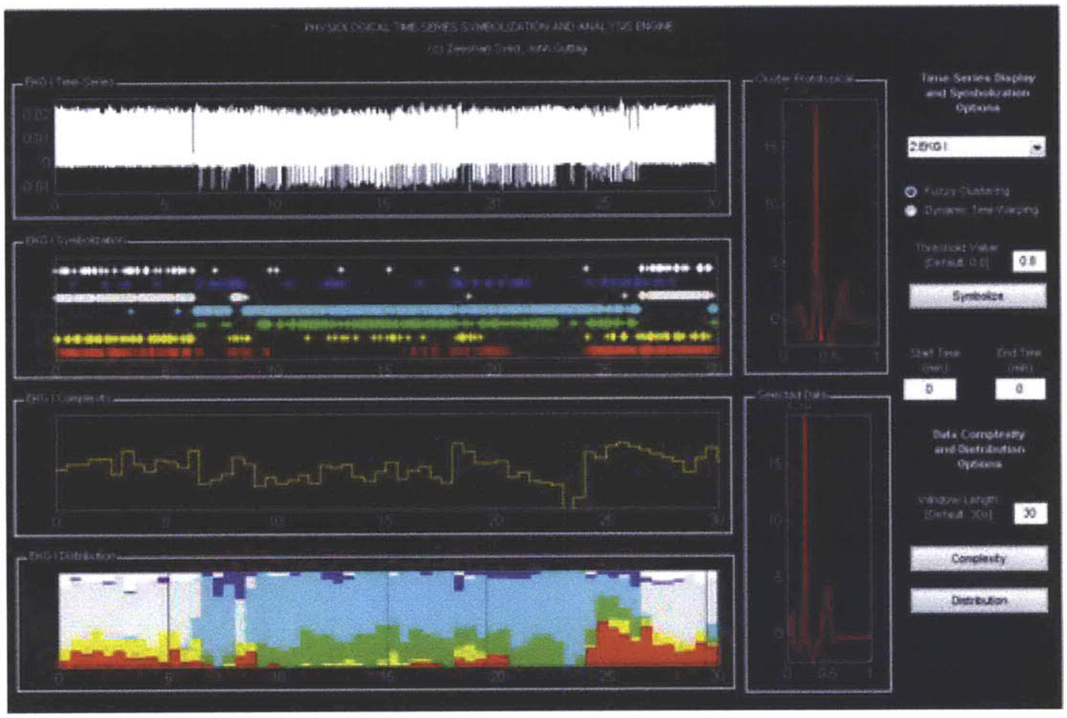
End Time (min) 30

◉ Fuzzy Clustering

○ Dynamic Time Warping

Threshold Value (Default 0.8) 0.8

Figure 5-1: Screenshot of compact symbolic display of long-term ECG.


PHYSIOLOGICAL TIME-SERIES SYMBOLIZATION AND ANALYSIS ENGINE
(c) Zeeshan Syed, John Guttag
EKG I Time-Series
EKG I Symbolization
EKG I Complexity
EKG I Distribution
Cluster Prototypical
Selected Data
Time Series Display and Symbolization Options
2:EKG I
Data Complexity and Distribution Options
Symbolize
Complexity
Distribution
0
5
10
15
20
25
30

● Fuzzy Clustering
○ Dynamic Time Warping

Threshold Value [Default: 0.8] 0.8

Start Time (min) 0

End Time (min) 0

Window Length [Default: 30s] 30

Figure 5-2: Screenshot of simultaneous ECG information display.

Figure 5-2 provides further information. In addition to showing the raw ECG signal (top panel) and the symbolic representation of this signal (second panel from top), it also shows the entropy of the symbolic sequence as a function of time (second panel from bottom), and how the distribution of different beat classes changes over time (bottom panel). Users can also click individual beats in the symbolic sequence to display the prototypical beat for the symbol, or the specific beat clicked (panels on right).

We use the entropy of the symbolic sequence as a proxy for the complexity of the underlying signal. As we showed in Section 4.2.2, changes in complexity may correspond to pathological conditions such as atrial fibrillation.

Figures 4-9 to 4-15, shown earlier in Section 4.2.2, were derived using our visualization tool and illustrate how our tool can be used to discover physiologically interesting activity. More generally, we expect clinicians to use our visualization tool to efficiently analyze long-term data (e.g., patient data between doctor visits) in a systematic manner. As a first step, this may involve noting the different symbols or classes of activity that took place over the course of long-term monitoring. Clinicians could then study how the distribution of the symbols changed over time, and potentially how the entropy or complexity of the signal evolved. Finally, clinicians could review interesting symbolic constructs, such as rhythms and frequently occurring patterns. These analyses would reveal information on what kind of activity occurred, when it occurred, and how it was organized. In Section 4.2.2 we showed how this information may correspond to important kinds of pathological activity.

In the remainder of this chapter, we now discuss our techniques for creating prototypical representations of physiological activity. The use of prototypes is an important supplement to the display of symbolic sequences, since it allows users to easily visualize a robust archetype of the activity corresponding to a symbol. Collectively, the display of symbols makes it easy to understand how different classes of activity change over time, while the prototype makes it easy to understand what each class represents.

5.2 Prototypical Representation of Biological Activity

Many biological signals exhibit quasi-periodicity, and this property encourages the use of a variety of aggregation techniques to create composite signals. Signal averaging is commonly employed to reduce additive noise and improve the quality of data. An associated advantage of this approach is that it can help address the problem of "information overload" by compressing the often overwhelming amounts of patient data collected during long-term monitoring.

Automated information fusion techniques and aggregation operations have been previously applied to a number of different fields. Ensemble averaging is frequently employed, and prototype construction in numerical, ordinal and categorical data has been explored using the plurality rule, medians, Sugeno integrals and ordinal weighted means [13]. The idea of weighted averaging [14] appears in various forms, using possibility theory [15], the minimum energy principle [16], signal-to-noise maximization with the Rayleigh quotient and generalized eigenvalue problem [17], and adaptive weight estimation [18]. In [19], a general discussion is provided on the application of averaging to quasi-periodic biological signals. Combination in this case proceeds by means of weighted averaging, where each observation influences the composite signal according to its weight and aggregation can be viewed as an optimization problem over the vector of weights using an objective function that measures the distance between the resulting prototype and each observation.

The works described above typically model all inter-period variations across the signals as noise. This includes variation due to time-skew, where the signal is not noise-corrupted, but may be variably delayed or advanced along its length. In this case, existing techniques for aggregation fail to make use of clean signal. We relax this view and attempt to separate noise from naturally occurring variations along the time axis in biological signals. This addresses the case of changes across signals resulting from misalignments of activity, rather than corruption of the data. To achieve this, we supplement previous aggregation algorithms with a mechanism to relate time-warped information across observations and combine data hierarchically. This ensures that the averaging process combines consistent physiological activity. We also make use of the differences in time-skew across observations to estimate an average length for each part of the signal. Under the time-warped model, we view the construction of prototypes as the merging of information along both the amplitude and time axes, i.e., the goal of our work is to produce representative signals where events have average amplitude and occur at average displacements from fiducial points.

5.2.1 Prototype Construction

As discussed in Section 3.2.3, the classic dynamic time-warping algorithm [83] produces the optimal alignment of two sequences $u[n]$ and $v[n]$, each of length N. For cardiovascular signals, $u[n]$ and $v[n]$ may be heartbeats. The optimal alignment can be denoted by $\phi = (\phi_u, \phi_v)$ and represents a mapping of length K between the samples

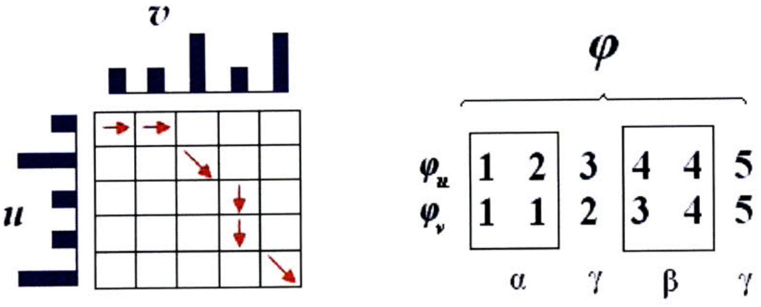


Figure 5-3: Decomposition of ϕ into α, β and γ segments. Horizontal runs along the DTW distance matrix correspond to α-segments, vertical ones to β-segments, and diagonal traversals to γ-segments.

of the two sequences, i.e.:

$$u[\phi_u(i)] \leftrightarrow v[\phi_v(i)] \; 1 \leq i \leq K \tag{5.1}$$

The process of aggregating $u[n]$ and $v[n]$ then proceeds by decomposing the optimal alignment ϕ into α, β and γ segments as shown in Figure 5-3. An α-segment corresponds to a locally maximal (i.e., non-extendible) segment of the alignment path such that multiple samples from $u[n]$ are aligned against a single sample from $v[n]$. Conversely, a β-segment comprises a locally maximal segment of the alignment path where multiple samples from $v[n]$ are aligned against $u[n]$. The remainder of ϕ corresponds to γ-segments.

Aggregation proceeds by iterating over all the decomposed segments in ϕ. Each α-segment of length l starting at position k (denoted by α_k) first undergoes a two-point average for $1 \leq n \leq \lfloor l/2 \rfloor$:

$$\alpha_k^{(1)}[n] = \frac{u[\phi_u(k+2n-1)] + u[\phi_u(k+2n)]}{4} \tag{5.2}$$

with:

$$\alpha_k^1[\lfloor\frac{l}{2}\rfloor + 1] = \frac{u[\phi_u(k+l-1)]}{2} \tag{5.3}$$

Information is then merged with v for $1 \leq n \leq \lfloor l/2 \rfloor + 1$ by:

$$\alpha_k^{(2)}[n] = \alpha_k(1)[n] + \frac{v[\phi_v(k)]}{2} \tag{5.4}$$

β-segments are handled in an analogous manner to yield $\beta_k^{(2)}[n]$. For γ-segments, there is no compacting or internal $\gamma_k^{(1)}[n]$ state. Instead, for $1 \leq n \leq l$:

$$\gamma_k^{(2)}[n] = \frac{u[\phi_n(k+n-1)] + v[\phi_n(k+n-1)]}{2} \tag{5.5}$$

The algorithm for aggregation cycles through all segments, and concatenates the respective $\alpha_k^{(2)}$, $\beta_k^{(2)}$ and $\gamma_k^{(2)}$ sub-signals together to yield an aggregate signal $x[n]$. For alternate α or β segments with even length (no distinction is drawn between the two cases), the last element corresponding to 5.4 is left out to ensure conservation of length.

It is critical that this pair-wise aggregation of observations yields a signal of the same length as the ones being combined. Conservation of length permits a hierarchical approach to compacting activity across multiple observations. We prove that this property is maintained by our algorithm for aggregation in Appendix A.

The combined effect of the halving associated with α_{odd}, β_{odd}, γ, α_{even} and β_{even} segments is that a total of $2N$ samples across the two signals combined is aggregated to yield a composite sequence of length N. The net effect of the pair-wise aggregation process is therefore to combine information in both amplitude and time, yielding a sequence that has the same length as the initial observations.

Figure 5-4 illustrates how a complete binary tree topology with depth $log(M)$ can be used to merge information in a uniform, unbiased manner. This contains $O(M)$ internal nodes, each corresponding to an $O(N^2)$ aggregation step associated with DTW-based pair-wise aggregation.

For very large observation sets, random sampling may be used to reduce M, or observations can be down sampled to obtain a quadratic speedup associated with the $O(N^2)$ factor.

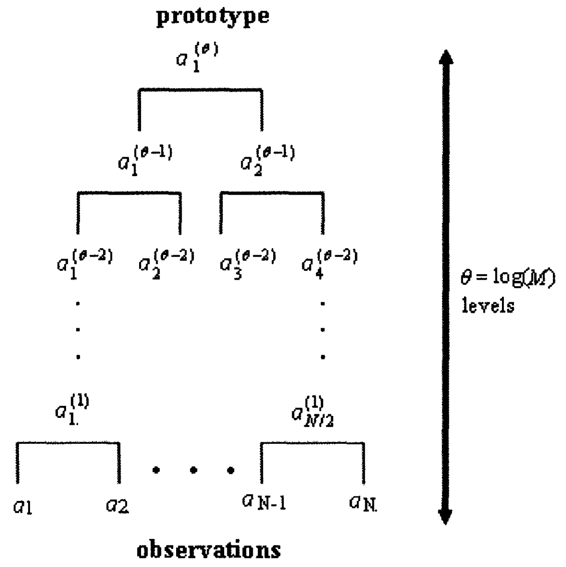


Figure 5-4: Hierarchical aggregation topology: The complete binary tree representation corresponds to a flat organization of observations that are aggregated in a breadth-first manner in pairs.

5.2.2 Evaluation

We used the synthetic ECG generator of [20] to evaluate the ability of our techniques to recover an original signal from observations corrupted by varying levels of noise and time-warping. Noise was specified directly to the generator in millivolts, while time-warping was expressed as the standard deviation of heart rate in terms of beats per minute. Since the number of observations generated was finite, the time-warping was not always centered around the original signal, i.e, the warped beats could on average be longer or shorter.

We tested our algorithm under various test conditions and calculated the root mean-square error (RMSE), maximal absolute difference (MD), signal-to-noise ratio (SNR), and dynamic time-warping cost (DTWC) between the aggregate prototype signals and the original signal. 100 observations were used, each corresponding to ECG sampled at 256 Hz with a mean heart rate of 80 beats per minute. The position of the R wave was used as the fiducial point.

The penalty under RMSE and SNR for a missing waveform can be half that for

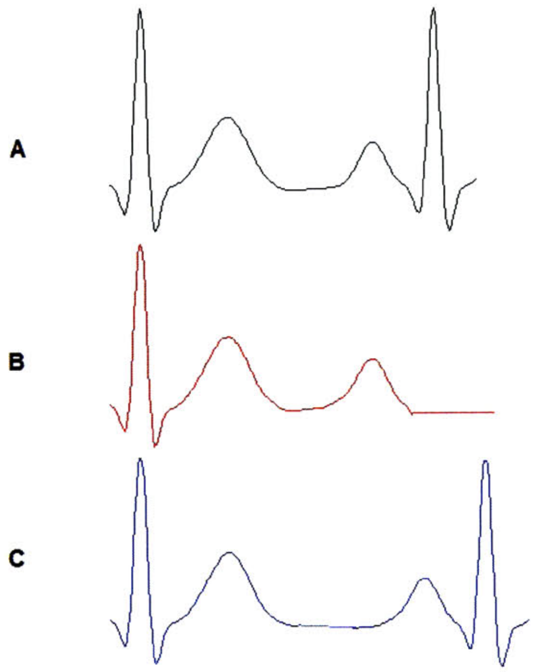


Figure 5-5: Example illustrating the limitations of RMSE and SNR as a measure of error in the presence of time-skew.

a slight change in the timing of that waveform. For example, consider the beats shown in Figure 5-5. RMSE and SNR would conclude that beats A and C were more dissimilar than beats A and B. This would occur because while comparing beats A and C, a slight delay in the R wave would lead to two mismatches, i.e., the R wave from beat A would correspond to low amplitudes parts of beat C, while the R wave from beat C would correspond to low amplitude parts of beat A. In contrast, even though beat B is missing the R wave altogether, RMSE and SNR would consider it more similar to beat A as there would only be one R wave mismatch (i.e., the R wave from beat A corresponding to low amplitude parts of beat B).

This phenomenon occurred in our experiments because of the finite sample issue discussed earlier. Similarly, the MD measure penalizes deletion as much as displacement. Since missing events altogether is typically more serious in the clinical setting than minor relocations, we consider the DTWC, which penalizes deletion more than displacement, to be a more useful measure of error.

The results of these experiments are shown in Figure 5-6. Each experiment is shown as a two dimensional grid, with noise increasing from top to bottom, and time-skew increasing from left to right. The color of each cell within the grid corresponds to the error. There are a total of eight grids in Figure 5-6, corresponding to the diffents experiments conducted. The left panel (i.e., the four grids in the left column) show error using each of the four measures between the ensemble average and the original signal. The right panel (i..e, the four grids in the right column) shows the corresponding errors for the prototype.

In the absence of time-warping, the ensemble provides a least squares optimal estimate of the original signal. This can be seen from the low error along the top of each grid on the left side of Figure 5-6. Even so, there is not much difference between the performance of the ensemble and that of the prototype in the absence of time-warping. Furthermore, with even small amounts of time-warping, the prototype provides a significant performance improvement. This can be seen from the greater increase in error (from top to bottom) in the grids on the left side of Figure 5-6 than on the right.

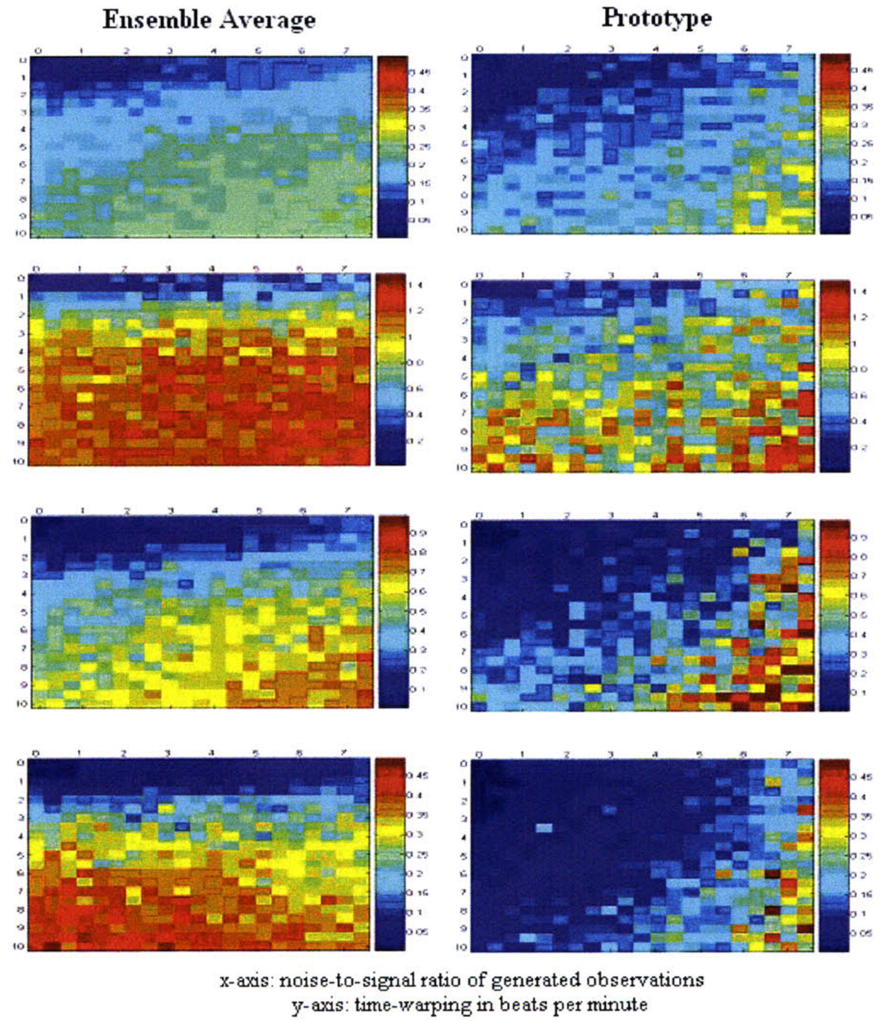


Figure 5-6: Synthetic ECG error: From top to bottom, the RMSE, MD, 1/SNR and DTWC of the ensemble average and prototype relative to the deterministic ECG signals are shown for different additive noise and warping levels (each cell corresponds to a different randomly generated set of observations). We use 1/SNR instead of SNR to display experiments corresponding to no noise (i.e., SNR= ∞ and 1/SNR= 0).

5.3 Summary

In this section, we presented our work on the visualization of information in large physiological datasets. We described a visualization tool that uses our symbolic framework to compactly present interesting activity in long-term physiological signals, and proposed algorithms to aggregate signals with noise and time-skew into prototypes.

Chapter 6

Conclusion and Future Work

We end with a summary of the major aspects of our work (Section 6.1), a review of our conclusions (Section 6.2), and a discussion of future research (Section 6.3).

6.1 Summary

In this thesis, we described several novel computational methods to analyze large amounts of continuous long-term physiological data. We focused largely on techniques to discover information that can be used to predict patients at high risk of adverse events, and applied these methods to cardiovascular datasets to risk stratify patients following non-ST-elevation acute coronary syndromes (NSTEACS).

6.1.1 New Concepts

We proposed two complementary areas of analysis.

Morphologic variability (MV) measures subtle micro-level changes in continuous signals, as a way to estimate instability in the underlying physiological system. These subtle variations have historically been considered to be noise. We show, however, that MV can provide a clinically useful estimate of instability in the underlying physiological system by identifying instabilities in the signal generated by that system. In the case of cardiac disease, we attribute these instabilities to the presence of unstable bifurcations in the myocardium causing increased variability in ECG morphology.

We also presented the notion of *symbolic analysis*. In contrast to our work on MV, this looks at macro-level information in signals by abstracting them into symbolic sequences and studying the resulting textual representations of the time series signals for interesting higher-level constructs (e.g., "words" within the textual representations that are associated with adverse outcomes). We explained how symbolization, which is used in many other disciplines, is useful within the context of physiological signals. More specifically, we demonstrated a symbolic framework that allows physiological datasets to be studied in a manner analogously to nucleotide data (albeit with different kinds of analyses).

We described two analyses for identifying high risk patients within a symbolic framework. We presented methods for discovering predictors of acute events by searching for approximate symbolic patterns that occur more often preceding events than one would expect by chance alone. We also introduced the idea of using comparative methods to relate long-term symbolic representations of time-series from different patients, and to identify "abnormal" patients as outliers in a population.

In addition to our work on morphologic variability and symbolic analysis, a third component of our work was the development of *visualization tools for long-term signals*. Our tools built upon our symbolic framework, and allow users to view continuous long-term signals as sequences of symbols while providing a mapping of each symbol to a prototypical waveform representation. This results in a large decrease in the number of data points that need to be visualized, makes it easier to see when changes occurred, and makes the data more readily interpretable.

As part of this work on visualization, we developed the idea of creating prototypical signals that aggregated both amplitude and time information. In contrast to conventional aggregation approaches, which average the amplitude of multiple observations, we proposed creating a signal where the duration of each physiological waveform was also averaged in time.

6.1.2 New Methods

In addition to proposing the concept of MV, we designed a system to measure MV, which addresses the challenge of finding subtle diagnostic variability in large amounts of noisy data with time-skew. Our algorithm uses a modified dynamic time-warping approach to compare variations in morphology between consecutive beats, and the Lomb-Scargle periodogram to identify a spectral signature for these variations that corresponds to high risk.

We also presented an efficient Max Min clustering-based algorithm for symbolization, and demonstrated that this transformation preserves useful clinical information while making the data easier to analyze. For example, we showed how different analyses on symbolic representations can be used to detect various kinds of clinically significant activity, e.g., searching for approximate repeated sequences finds ventricular bigeminy and trigeminy; searching for statistically overrepresented patterns reveals tachyarrhythmias; and locating high entropy periods detects atrial fibrillation. Our algorithms also uncovered kinds of complex activity that often go unnoticed in clinical practice, e.g., atrial ectopic rhythms.

We approached the problem of pattern discovery as a significance and classification problem, and used the ideas of locality sensitive hashing (LSH), multi-level Gibbs sampling, and sequential statistics to make the search for relevant activity more efficient. We also developed the symbolic mismatch method to compare long-term symbolic representations of time-series from different patients, by reducing the time-series for each patient to prototypical segments and measuring the probability-weighted mismatch of these segments across patients to assess similarity. This approach was used to partition patients into groups with similar risk profiles.

6.1.3 New Clinical Results

We developed MV using data from the DISPERSE2 TIMI33 study of approximately 800 patients. When evaluated on over 4,500 patients from the MERLIN TIMI36 study, we found that high MV was associated with cardiovascular death and sudden cardiac death post-NSTEACS. These results were consistent even after adjusting for clinical characteristics, biomarker data, and medications. Our data show that information in MV may be independent of information provided by the other risk variables, and in particular, MV is a better predictor of death than other long-term ECG-based metrics. MV also has great value in identifying patients who are missed by echocardiography. The adjusted hazard ratios in patients with LVEF$\geq$40% were 2.93 ($p<0.001$) for cardiovascular death and 2.27 ($p=0.013$) for sudden cardiac death.

We demonstrated the utility of our symbolic analysis methods to discover predictors of acute events, both for detecting markers associated with long-term risk and for markers associated with imminent acute events (e.g., sudden cardiac death). In a small study on patients from the Physionet Sudden Cardiac Death database, who experienced sudden cardiac death, our algorithms correctly identified 70% of the patients who died while classifying none of the normal individuals and only 8% of the patients with supraventricular arrhythmias as being as risk.

We evaluated our comparative methods on data from roughly 800 patients in the DISPERSE2 TIMI33 study. We used our methods to partition patients with cardiovascular disease into groups using their ECG signals. We found that different groups of patients exhibit a varying risk of adverse outcomes. One group, with a particular set of time-series characteristics, shows a 23 fold increased risk of death, while another exhibits a 5 fold increased risk of future heart attacks.

While most of our clinical results address the problem of cardiovascular risk stratification within adult populations, we also showed that both morphologic variability and symbolic analysis are more broadly useful. In particular, we demonstrated how these tools can help quantify the effects of pharmaceutical treatment, and may also have value in applications such as fetal risk stratification for cerebral palsy.

6.2 Conclusions

The practice of medicine can be simplified into different steps, e.g., evaluating patients, choosing interventions and performing interventions. Perhaps the weakest link in this loop is the evaluation of patients. This is one reason that the existing practice of medicine is largely curative rather than preventative.

Computational techniques and tools can play a critical role in addressing this problem. They can improve our ability to understand what constitutes risk and help us develop real-time methods for identifying high risk patients. We can achieve both these goals without raising cost or burdening caregivers or patients.

Our work on analyzing ECG data represents exactly such a scenario. Our methods are able to identify high risk patients using data that is already routinely collected in hospitals and ambulatory monitoring. We are therefore potentially able to make a positive impact without the need for any new hardware or patient maneuvers. Since our tools are fully automated, we do not expect a significant increase in clinician time.

The success of our work on ECG signals also helps make an important point: there is often much information, even in traditional signals like the ECG that have been around since the late nineteenth century, that can be used to make a significant difference. This is particularly true of long-term recordings, which are challenging to analyze with the human eye and have only recently become widely available because of advances in recording technologies. The challenge, however, is extracting the information in these signals, and it is in this context that computational tools to analyze physiological signals can make a useful contribution.

6.3 Future Work

In the near term, there is a need to reproduce our results on other datasets, so that our risk metrics can be incorporated into clinical practice. While our studies were prospectively designed, the data on which they were conducted was part of concluded drug trials. We hope to carry out similar analyses in a prospective trial, on larger patient populations. We also hope to study patients with ST-elevation MI in addition to NSTEACS, so that we may be able to more completely evaluate our tools for risk stratification post-ACS.

A different extension of our work would be to study our tools on data from patients who have no prior history of coronary heart disease, and evaluate their utility in risk stratifying patients for *de novo* events. While we believe that techniques such as MV may have value in this context, this theory needs to be rigorously tested.

There is also a need to validate our pathophysiological theory for morphologic variability. This will require animal studies and high resolution imaging to confirm that variability in the shape of the ECG signals is due to unstable bifurcations.

For morphologic variability, another important area of work is to reduce the amount of data needed to measure MV. Presently, our algorithm for measuring MV requires roughly ten hours worth of data. This restricts MV to use on patients who are hospitalized or subject to ambulatory monitoring. Being able to reduce the amount of data needed to measure MV would facilitate using this technique to identify individuals within the general population who are at high risk of adverse cardiovascular outcomes.

We also hope to extend our work to signals other than the ECG. In the near future, this may correspond to other easily segmentable signals (e.g., blood pressure and respiration). Using related computational methods to analyze signals from the brain represent another potentially exciting area of research. However, we believe that analyzing signals emanating from the brain is likely to require methods that are quite different from those we have used thus far.

We would also like to address a limitation of our present approach, i.e., both morphologic variability and symbolic analysis study time-series signals as direct observations of the underlying physiological system. It does not attempt to relate changes in these observations to changes in the system being observed (i.e., the heart). We believe that there is a need to extend our work to include a process of deconvolution, e.g., through Hidden Markov Models (HMMs) or blind system identification methods, that allows one to study the internal state of physiological systems rather than operate at the level of their stochastic outputs. Such methods might, for example, reveal how the underlying functional states change from baseline periods to periods preceding acute events such as fatal arrhythmias.

Another potential area of future work is the study of multi-signal trends. In medicine, there are often tens of relevant variables to consider; far more than a physician is able to assimilate. We believe that our work on symbolization may help address this situation. In particular, we would like to develop computational methods for deducing clinically significant interactions among multiple long-term time-series. The search space for these multi-modal interactions is, of course, immense. It can be reduced using some of the techniques discussed above, but that will not be sufficient. One approach we consider promising is to produce synthetic signals that combine multiple time-series, e.g., signals that combine the electrical activity measured by an ECG with the pressure waveform associated with blood pressure. We also believe that truncation approaches for measuring mutual information may be successful.

Finally, we also believe there is a need to test our visualization tools more rigorously. Presently, these tools have been restricted to experimental use. We hope to make them publicly available, so that a more comprehensive understanding of the value of these tools can be developed.

In summary, this thesis represents an initial attempt to develop computaitonal methods to analyze large multi-patient, multi-signal physiological datasets. We expect this area to grow further, and believe it may have a valuable role in helping accelerate the progress of medicine.

Appendix A

Proof of Conservation of Length

Theorem: For two inputs $u[n]$ and $v[n]$ of length N, the $x[n]$ produced by the aggregation algorithm in Section 5.2 is also of length N.

Proof: The total length of the aggregated sequence $x[n]$ can be found by distinguishing between α and β segments of even or odd length. α_{odd} and β_{odd} segments of odd length l merge together $l+1$ samples across both signals (l samples along one signal, and one sample along the other) to yield $\alpha^{(2)}$ or $\beta^{(2)}$ sub-signals of length $\lfloor l/2 \rfloor + 1$. Since l is odd:

$$\lfloor \frac{l}{2} \rfloor + 1 = \frac{l+1}{2} \tag{A.1}$$

Each α_{odd} or β_{odd} segment therefore reduces the distinct samples along it (i.e., the samples from both signals that comprise these segments) by half. A similar effect can be seen for γ-segments, each of which compact $2l$ distinct samples into a $\gamma^{(2)}$ signal of length l. What remains to be proven is that α_{even} and β_{even} segments also compact the number of distinct samples along them by half.

In the case of α_{even} and β_{even} segments, the corresponding lengths of $\alpha^{(2)}$ and $\beta^{(2)}$ are alternately $\lfloor l/2 \rfloor+1$ and $\lfloor l/2 \rfloor$. This follows from the fact that, as described earlier, for alternate α or β segments with even length, the last element corresponding to 5.4 is left out. The resulting effect due to this approach is that every odd occurrence of an α_{even} or β_{even} segment has a length exceeding $(l+1)/2$ by half a sample, while every even occurrence has a length that is less by the same amount, which counterbalances the earlier excess. In other words, the net effect of having an even number of α_{even} and β_{even} segments is that cumulatively, these segments compact samples along both signals by half.

We can prove the existence of an even number of α_{even} and β_{even} segments as follows. Each DTW alignment path starting at $(1,1)$ and ending at (N,N) can be described through $2N1$ movements in the $\rightarrow$ or $\downarrow$ direction. Denoting the m-th α, β, and γ segments by α_m, β_m, and γ_m and length by the operator $L(.)$, it is easy to verify that:

$$2N-1=\sum_{m_\alpha} L(\alpha_{m_\alpha})+\sum_{m_\beta} L(\beta_{m_\beta})+2\sum_{m_\gamma} L(\gamma_{m_\gamma})+m_\alpha+m_\beta-1 \tag{A.2}$$

Replacing the summation terms by w_α, w_β and w_γ:

$$2N-1=w_\alpha+w_\beta+2w_\gamma+m_\alpha+m_\beta-1 \tag{A.3}$$

Distinguishing between even (e) and odd (o) length terms this can be rewritten as:

$$2N=w_{\alpha,e}+w_{\alpha,o}+w_{\beta,e}+w_{\beta,o}+2w_\gamma+m_{\alpha,e}+m_{\alpha,o}+m_{\beta,e}+m_{\beta,o} \tag{A.4}$$

The terms corresponding to $2w_{gamma}$, $w_{\alpha,e}$ and $w_{\beta,e}$ must always be even. Removing them:

$$2N'=w_{\alpha,o}+w_{\beta,o}+m_{\alpha,e}+m_{\alpha,o}+m_{\beta,e}+m_{\beta,o} \tag{A.5}$$

$w_{\alpha,o}$ can only be odd if an odd number of α_{odd}-segments are present, i.e., $m_{\alpha,o}$ is odd. Similarly, the condition for $w_{\beta,o}$ to be odd is that $m_{\beta,o}$ is odd. From this it follows that:

$$w_{\alpha,o} + w_{\beta,o}\,(mod2) = m_{\alpha,o} + m_{\beta,0}\,(mod2) \tag{A.6}$$

This means that:

$$2N'' = m_{\alpha,e} + m_{\beta,e} \tag{A.7}$$

An even number of α_{even} and β_{even} segments must therefore exist.

Bibliography

[1] Goldberger JJ, Cain ME, Hohnloser SH, Kadish AH, Knight BP, Lauer MS, Maron BJ, Page RL, Passman RS, Siscovick D, Stevenson WG, Zipes DP, 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: a scientific statement from the American Heart Association Council on Clinical Cardiology Committee on Electrocardiography and Arrhythmias and Council on Epidemiology and Prevention, Circulation, 2008.

[2] Elliot VS, Blood pressure readings often unreliable, Am Med News, 2007.

[3] World Health Organization, Cardiovascular diseases, Fact Sheet, 2005.

[4] Hotelling H, Analysis of a complex of statistical variables into principal components, J Educ Psych, 1933.

[5] Eckberg D, Sympathovagal balance: a critical appraisal, Circulation, 1997.

[6] Malliani A, Pagani M, Montano N, Sympathovagal balance: a reappraisal, Circulation, 1998.

[7] Gomez R, Romero R, Ghezzi F, Yoon BH, Mazor M, Berry SM, The fetal inflammatory response syndrome, Am J Obstet Gynecol, 1998.

[8] Shimoya K, Matsuzaki N, Taniguchi T, Okada T, Saji F, Murata Y, Interleukin-8 level in maternal serum as a marker for screening of histological chorioamnionitis at term, Int J Gynaecol Obstet, 1997.

[9] Sorensen K, Brodbeck U, Norgaard-Pedersen B, Determination of neuron specific enolase in amniotic fluid and maternal serum for the prenatal diagnosis of fetal neural tube defects, Clin Chim Acta, 1987.

[10] Wu Y, Colford JM, Chorioamnionitis as a risk factor for cerebral palsy, JAMA, 2000.

[11] Zupan V, Gonzalez P, Lacaze-Masmonteil T, Boithias C, d'Allest AM, Dehan M, Gabilan JC, Periventricular leukomalacia: risk factors revisited, Dev Med Child Neurol, 1996.

[12] Smulian JC, Vintzileos AM, Lai YL, Santiago J, Shen-Schwarz S, Campbell WA, Maternal chorioamnionitis and umbilical vein interleukin-6 levels for identifying early neonatal sepsis, J Mat Fet Med, 1999.

[13] Domingo-Ferrer J, Torra V, Median-based aggregation operators for prototype construction in ordinal scales, Int J Intel Sys, 2003.

[14] Yager R, On mean type aggregation, IEEE Trans Syst Man Cyber, 1996.

[15] Dubois D, Prade H, Possibility theory in information fusion, Int Conf Inf Fus, 2000.

[16] Fan Z, Wang T, Weighted averaging method for evoked potentials: determination of weighted coefficients, IEEE EMBS , 1992.

[17] Davila C, Mobin M, Weighted averaging of evoked potentials, IEEE Trans Biomed Eng, 1992.

[18] Bataillou E, Thierry E, Rix H, Meste O, Weighted averaging using adaptive estimation of the weights, signal processing, 1995.

[19] Leski J, Robust weighted averaging, IEEE Tran Biomed End, 2002.

[20] McSharry P, Clifford G, Tarassenko L, Smith L, A dynamic model for generating synthetic electrocardiogram signals, IEEE Trans Biomed Eng, 2003.

[21] Lin J, Keogh E, Lonardi S, Chiu B, A symbolic representation of time series, with implications for streaming algorithms, SIGKDD, 2003.

[22] Crooks GE, Hon G, Chandonia JM, Brenner SE, WebLogo: A sequence logo generator, Genome Res, 2004.

[23] Sneath PHA, Sokal RR, Numerical taxonomy, Freeman, 1973.

[24] Duda R, Hart P, Pattern classification, Wiley-Interscience, 2000.

[25] Chang S, Cheng F, Hsu W, Wu G, Fast algorithm for point pattern matching: invariant to translations, rotations and scale changes, Pattern Recognition, 1997.

[26] Cohen WW, Richman J, Learning to match and cluster large high-dimensional data sets for data integration, SIGKDD, 2002.

[27] Newby LK, Bhapkar MV, White HD, Topol EJ, Dougherty FC, Harrington RA, Smith MC, Asarch LF, Califf RM, Predictors of 90-day outcome in patients stabilized after acute coronary syndromes, Eur Heart J, 2003.

[28] Bland M, Altman DG, Multiple significance tests: the Bonferroni method, BMJ, 1995.

[29] Phatarfod RM, Sudbury A, A simple sequential Wilcoxon test, Aus J Stat, 30:93-106, 1988.

[30] Brazma A, Jonassen I, Eidhammer I, Gilber D, Approaches to the automatic discovery of patterns in biosequences, J Comp Biol, 1998.

[31] Buhler J, Efficient large-scale sequence comparison by locality-sensitive hashing, Bioinformatics, 2001.

[32] Buhler J, Tompa M, Finding motifs using random projections, J Comp Biol, 2002.

[33] Lawrence CE, Altschul SF, Boguski, Liu JS, Neuwald AF, Wootton JC, Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment, Science, 1993.

[34] Wilcoxon F, Individual comparisons by ranking methods, Biometrics Bulletin, 1945.

[35] Grundy WN, Bailey TL, Elkan CP, Baker ME, Meta-MEME: motif-based hidden Markov models of protein families, Comp Appl Biosci, 1997.

[36] Haveliwala TH, Gionis A, Indyk P, Scalable techniques for clustering the web, WebDB Workshop, 2000.

[37] Hochbaum DS, Shmoys DB, A best possible heuristic for the k-center problem, Math Op Res, 1985.

[38] Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM, Systematic determination of genetic network architecture, Nat Genetics, 1999.

[39] Liu S, Brutlag DL, Liu JS, BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes, Pac Symp Biocomp, 2001.

[40] Lv Q, Josephson W, Wang Z, Charikar M, Li K, Multi-probe LSH: efficient indexing for high-dimensional similarity search, VLDB, 2007.

[41] Sinha S, Tompa M, YMF: a program for discovery of novel transcription factor binding sites by statistical overrepresentation, Nucl Acids Res, 2003.

[42] Batzoglou S, Pachter L, Mesirov JP, Berger B, Lander ES, Human and mouse gene structure: comparative analysis and its application to exon prediction, Genome Res, 2000.

[43] Delcher AL, Kasif S, Fleischmann RD, Peterson J, White O, Salzberg SL, Alignment of whole genomes. Nucl Acids Res, 1999.

[44] Indyk P, Motwani R, Approximate nearest neighbors: towards removing the curse of dimensionality, Sym Theor Comput, 1998.

[45] Lehmann EL, Nonparametrics: statistical methods based on ranks, McGraw-Hill, 1975.

[46] Reynolds MR, A sequential signed-rank test for symmetry, Ann Stat, 1975.

[47] Hollander M, Wolfe DA, Nonparametric statistical methods, John Wiley and Sons, 1999.

[48] Gibbons JD, Nonparametric statistical inference, Marcel Dekker, 1985.

[49] Gert T, Marchal K, Lescot M, Rombauts S, De Moor B, Rouze P, Moreau Y, A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes, J Comp Biol, 2003.

[50] Stormo G, Hartzell G, Identifying protein-binding sites from unaligned DNA fragments, PNAS, 1989.

[51] Jones N, Pevzner P, An introduction to bioinformatics algorithms. The MIT Press, 2004.

[52] Kellis M, Patterson N, Endrizzi M, Birren B, Lander E, Sequencing and comparison of yeast species to identify genes and regulatory elements, Nature, 2003.

[53] Bailey T, Eklan C, The value of prior knowledge in discovery motifs with MEME, ICISMB, 1995.

[54] Kojadinovic I, Relevance measures for subset variable selection in regression problems based on k-additive mutual information, Comput Stat Data Anal, 2004.

[55] Abramson N, Information theory and coding, McGraw Hill, 1963.

[56] Jennings D, Amabile T, Ross L, Information covariation assessments: data-based versus theory-based judgements, Judgement under uncertainty: heuristics and biases, Cambridge University Press, 1982.

[57] Baumert M, Baier V, Truebner S, Schirdewan A, Voss A, Short and long term joint symbolic dynamics of heart rate and blood pressure in dilated cardiomyopathy, IEEE Trans Bio Eng, 2005.

[58] Altschul S, Gish W, Miller W, Myers E, Lipman D, Basic local alignment search tool, J Mol Biol, 1990.

[59] Landau G, Schmidt J, Sokol D, An algorithm for approximate tandem repeats, J Mol Biol, 2001.

[60] Hamming R, Error-detecting and error-checking codes, Bell System Technical Journal, 1950.

[61] Syed Z, Guttag J, Prototypical biological signals, ICASSP, 2007.

[62] Chen G, Wei Q, Zhang H, Discovering similar time-series patterns with fuzzy clustering and DTW methods, NAFIPS, 2001.

[63] Keogh E, Pazzani M, Scaling up dynamic time warping for data mining applications, SIGKDD, 2000.

[64] Gonzalez T, Clustering to minimize the maximum intercluster distance, Theor Comp Sci, 38; 1985.

[65] Frau-Cuesta D, Perez-Cortes J, Andreu-Garcia G, Clustering of electrocardiograph signals in computer-aided Holter analysis, Comp Methods Programs Biomed, 2003.

[66] Braunwald E, Zipes D, Libby P, Heart disease: a textbook of cardiovascular medicine, WB Saunders Co, 2001.

[67] Daw C, Finney C, Tracy E, A review of symbolic analysis of experimental data, Rev Sci Instrum, 2003.

[68] Sung PP, Risk stratification by analysis of electrocardiographic morphology following acute coronary syndromes, MIT EECS Thesis, 2008.

[69] Syed Z, MIT automated auscultation system, MIT EECS Thesis, 2003.

[70] Kopec D, Kabir MH, Reinharth D, Rothschild O, Castiglione JA, Human errors in medical practice: systematic classification and reduction with automated information systems, J Med Syst, 2003.

[71] Martich GD, Waldmann CS, Imhoff M, Clinical informatics in critical care, J Intensive Care Med, 2004.

[72] American Heart Association, Heart Disease and Stroke Statistics – 2009 Update, 2009.

[73] Feder BJ, Defibrillators Are Lifesaver, but Risks Give Pause, New York Times, 2008.

[74] DeChazal P, ODwyer M, Reilly R, Automatic classification of heartbeats using ECG morphology and heartbeat interval features, IEEE Trans Biomed Eng, 2004.

[75] Ben-Haim SA, Becker B, Edoute Y, Kochanovski M, Azaria O, Kaplinksy E, Palti Y, Beat-to-beat electrocardiographic morphology variation in healed myocardial infarction, Am J Cardiol, 1991.

[76] DeChazal P, ODwyer M, Reilly R, Automatic classification of heartbeats using ECG morphology and heartbeat interval features, IEEE Trans Biomed Eng, 2004.

[77] Donoho D, Denoising by soft-thresholding, IEEE Trans Inf Theory, 2005.

[78] El-Sherif N, Hope RR, Scherlag BJ, Lazara R, Reentrant ventricular arrhythmias in the late myocardial infarction period, Circulation, 1977.

[79] Hamilton PS, Tompkins WJ, Quantitative investigation of QRS detection rules using the MIT/BIH arrhythmia database, IEEE Trans Biomed Eng, 1986.

[80] Hamilton PS, Open source ECG analysis software documentation, EPLimited, 2002.

[81] Josephson ME, Wit AL, Fractionated electrical activity and continuous electrical activity. Fact or artifact? Circulation, 1984.

[82] Li Q, Mark RG, Clifford GD, Robust heart rate estimate fusion using signal quality indices and a Kalman filter, Physiol Meas, 2008.

[83] Myers C, Rabiner L, Rosenberg A, Performance tradeoffs in dynamic time-warping algorithms for isolated word recognition, IEEE Trans Acoust, 1980.

[84] Ohman E, Granger C, Harrington R, Lee K, Risk stratification and therapeutic decision-making in acute coronary syndromes, JAMA, 2000.

[85] Rabiner L, Considerations in dynamic time-warping algorithms for discrete word recognition, IEEE Trans Signal Process, 1978.

[86] Zong W, Moody GB, Jiang D, A robust open-source algorithm to detect onset and duration of QRS complexes, Comput Cardiol,2003.

[87] Syed Z, Guttag J, Stultz C, bClustering and symbolic analysis of cardiovascular signals: discovery and visualization of medically relevant patterns in long-term data with limited prior knowledge, EURASIP J Adv Signal Process, 2007.

[88] Cannon CP, Husted S, Harrington RA, Scirica BM, Emanuelsson H, Peters G, Storey RF, 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, J Am Coll Cardiol, 2007.

[89] Lomb NR, Least-squares frequency analysis of unequally spaced data, Astrophys Space Sci, 1976.

[90] Bauer A, Kantelhardt JW, Barthel P, Schneider R, Ing D, Makikallio T, Ulm K, Hnatkova K, Shomig A, Huikuri H, Bunde A, Malik M, Schmidt G, Deceleration capacity of heart rate as a predictor of mortality after myocardial infarction: cohort study, Lancet, 2006.

[91] Cox DR, Oakes D, Analysis of survival data, Chapman and Hall. 1984.

[92] Meyers L, Gamst G, Guarino A, Applied multvariate research: design and interpretation, Sage Publications, 2005.

[93] Malik M, Heart rate variability: standards of measurement, physiological interpretation, and clinical use, Circulation, 1996.

[94] Morrow DA, Scirica BM, Prokopczuk E, Murphy SA, Budaj A, Varshavsky S, Wolff AA, Skene A, McCabe CH, Braunwald E, Effects of ranolazine on recurrent cardiovascular events in patients with non-ST-elevation acute coronary syndromes: The MERLIN-TIMI 36 randomized trial, JAMA, 2007.

[95] Lilly LS, Pathophysiology of heart disease, Lippincott Williams and Wilkins, 2003.

[96] Cannon CP, Braunwald E, McCabe CH, Rader DJ, Roulea JL, Belder R, Joyal SV, Hill KA, Pfeffer MA, Skene AM, Intensive versus moderate lipid lowering with statins after acute coronary syndromes, N Engl J Med, 2004.

[97] Anderson JL, Adams CD, Antman EM, Bridges CR, Califf RM, Casey DE Jr, Chavey WE 2nd, Fesmire FM, Hochman JS, Levin TN, Lincoff AM, Peterson ED, Theroux P, Wenger NK, Wright RS, Smith SC Jr, Jacobs AK, Adams CD, Anderson JL, Antman EM, Halperin JL, Hunt SA, Krumholz HM, Kushner FG, Lytle BW, Nishimura R, Ornato JP, Page RL, Riegel B, ACC/AHA 2007 guidelines for the management of patients with unstable angina/non-ST-Elevation myocardial infarction: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (Writing Committee to Revise the 2002 Guidelines for the Management of Patients With Unstable Angina/Non-ST-Elevation Myocardial Infarction) developed in collaboration with the American College of Emergency Physicians, the Society for Cardiovascular Angiography and Interventions, and the Society of Thoracic Surgeons endorsed by the American Association of Cardiovascular and Pulmonary Rehabilitation and the Society for Academic Emergency Medicine, J Am Coll Cardiol, 2007.

[98] Billman GE, Schwartz PJ, Stone HL, Baroreceptor reflex control of heart rate: a predictor of sudden cardiac death, Circulation, 1982.

[99] Schwartz PJ, Vanoli E, Stramba-Badiale M, De Ferrari GM, Billman GE, Foreman RD, AAutonomic mechanisms and sudden death. New insights from analysis of baroreceptor reflexes in conscious dogs with and without a myocardial infarction, Circulation, 1988.

[100] Chess GF, Tam RMK, Calaresu FR, Influence of cardiac neural inputs on rhythmic variations of heart period in the cat, Am J Physiol, 1975.

[101] Tsuji H, Venditti FJ, Manders ES, Evans JC, Larson MG, Feldman CJ, Levy D. Reduced heart rate variability and mortality risk in an elderly cohort, Circulation, 1994.

[102] Zwanziger J, Hall WJ, Dick AW, Zhao H, Mushlin AI, Hahn RM, Wang H, Andrews ML, Mooney C, Wang H, Moss AJ, The cost effectiveness of implantable cardioverter-defibrillators: results from the Multicenter Automatic Defibrillator Implantation Trial (MADIT)-II, J Am Coll Cardiol, 2006.

[103] Caruso A, Marcus F, Hahn E, Hartz V, Mason J, Predictors of arrhythmic death and cardaic arrest in the ESVEM trial, Circulation, 1997.

[104] Barthel P, Schneider R, Bauer A, Ulm K, Schmitt C, Schomig A, Schmidt G, Risk stratification after acute myocardial infarction by heart rate turbulence, Circulation, 2003.

[105] Nissen SE, Tuzcu EM, Schoenhagen P, Brown BG, Ganz P, Vogel RA, Crowe T, Howard G, Cooper CJ, Brodie B, Grines CL, DeMaria AN, Effect of intensive compared with moderate lipid-lowering therapy on progression of coronary atherosclerosis: a randomized controlled trial, JAMA, 2004.

[106] Antman EM, Cohen M, Bernink PJ, McCabe CH, Horacek T, Papuchis G, Mautner B, Corbalan R, Radley D, Braunwald E, The TIMI risk score for unstable

angina/non-ST elevation MI: a method for prognostication and therapeutic decision making, JAMA, 2000.

[107] Morrow DA, Antman EM, Charlesworth A, Cairns R, Murphy SA, de Lemos JA, Giugliano RP, McCabe CH, Braunwald E, TIMI risk score for ST-elevation myocardial infarction: a convenient bedside clinical score for risk assessment at presentation, Circulation, 2000.

[108] Morrow DA, Antman EM, Parsons L, de Lemos JA, Cannon CP, Giugliano RP, McCabe CH, Barron HV, Braunwald E, Application of the TIMI risk score for ST-elevation MI in the National Registry of Myocardial Infarction 3, JAMA, 2001.

[109] Gumina R, Wright R, Kopeckya S, Millera W, William B, Reeder G, Murphy J, Strong predictive value of TIMI risk score analysis for in-hospital and long-term survival of patients with right ventricular infarction, Eur Heart J, 2002.

[110] Bigger JT, Fleiss JL, Rolnitzky LM, Steinman RC, The ability of several short-term measures of RR variability to predict mortality after myocardial infarction, Circulation, 1993.

[111] Kleiger RE, Miller JP, and Bigger JT, Decreased heart rate variability and its association with increased mortality after acute myocardial infarction, Am J Cardiol, 1987.

[112] Syed Z, Scirica BM, Stultz CM, Guttag, JV, Risk-stratification following acute coronary syndromes using a novel electrocardiographic technique to measure variability in morphology, Comput Cardiol, 2008.

[113] Syed Z, Scirica BM, Mohanavelu S, Sung P, Michelson EL, Cannon CP, Stone PH, Stultz CM, Guttag JV, Relation of death within 90 days of non-ST-elevation acute coronary syndromes to variability in electrocardiographic morphology, Am J Cardiol, 2009.

[114] Syed Z, Sung P, Scirica BM, Morrow DA, Stultz CM, Guttag JV, Spectral energy of ECG morphologic differences to predict death, Cardiovasc Eng, 2009.

[115] Smith JM, Clancy EA, Valeri CR, Ruskin JN, Cohen RJ, Electrical alternans and cardiac electrical instability, Circulation, 1988.

[116] Hanley JA, McNeil BJ, The Meaning and Use of the area under a Receiver Operating Characteristic (ROC) curve, Radiology, 1982.

[117] Granger CB, Goldberg RJ, Dabbous O, Pieper KS, Eagle KA, Cannon CP, Van De Werf F, Avezum A, Goodman SG, Flather MD, Fox KA, Predictors of hospital mortality in the global registry of acute coronary events, Arch Intern Med, 2003.

[118] Boersma E, Pieper KS, Steyerberg EW, Wilcox RG, Chang WC, Lee KL, Akkerhuis KM, Harrington RA, Deckers JW, Armstrong PW, Lincoff AM, Califf RM, Topol EJ, Simoons ML, Predictors of outcome in patients with acute coronary syndromes without persistent ST-segment elevation : results From an international trial of 9461 patients, Circulation, 2000.

[119] Krumholz HM, Douglas PS, Goldman L, Waksmonski C, Clinical utility of transthoracic two-dimensional and Doppler echocardiography, J Am Coll Cardiol, 1994.

[120] Moss AJ, Zareba W, Hall WJ, Klein H, Wilber DJ, Cannom DS, Daubert JP, Higgins SL, Brown MW, Andrews ML, Prophylactic implantation of a defibrillator in patients with myocardial infarction and reduced ejection fraction, N Engl J Med, 2002.

[121] Shoeb A, Edwards H, Connolly J, Bourgeois B, Treves S, Guttag J, Patient-specific seizure onset detection, Epilepsy Behav, 2004.

[122] Agarwal R, Gotman J, Long-term EEG compression for intensive-care settings, IEEE Eng Med Biol, 2001.

[123] Gustafson D, Kessel W, Fuzzy clustering with a fuzzy covariance matrix, IEEE CDC, 1978.