Learning Better Clinical Risk Models

by

Alexander Van Esbroeck

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Computer Science and Engineering)
in The University of Michigan
2015

Doctoral Committee:

Professor Satinder Singh Baveja, Co-Chair
Assistant Professor Zeeshan Syed, Co-Chair
Assistant Professor Mariel Lavieri
Assistant Professor Honglak Lee

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF ABBREVIATIONS

CVD Cardiovascular Death

SCD Sudden Cardiac Death

ACS Acute Coronary Syndrome

LVEF Left Ventricular Ejection Fraction

ICD Implantable Cardiac Defibrillator

ECG Electrocardiogram

TRS TIMI Risk Score

EEG Electroencephalogram

ICU Intensive care unit

AUC Area under the ROC curve

HR Hazard Ratio

IDI Integrated discrimination improvement

NRI Net reclassification improvement

MAR Missing at random

NMAR Not missing at random

MCAR Missing completely at random

LDA Latent Dirichlet Allocation

CHAPTER I

Introduction

Clinical risk stratification is the determination of a patient's risk of suffering a particular outcome. Models that combine a variety of risk factors, such as demographics, medical history, and clinical and lab tests, are used to predict patient risk of adverse outcomes in many areas of clinical practice, including cardiac [1, 28, 75] and critical care [44, 2, 16] patients, as well as to predict readmission after discharge from the hospital [35]. The estimates of patient risk that these models generate are used to make critical decisions, including choosing treatment plans for patients, allocating limited resources in hospitals, and evaluating the performance of healthcare providers. More accurate predictions of risk, either by incorporating more discriminative risk factors into these models, or through the learning of more accurate models, result in improvements in each of these processes across many clinical areas.

More accurate prediction of adverse outcomes lets clinicians match patients to more effective treatments. Underestimation of a patient's risk causes patients who need more intensive treatment to go without it, while overestimation leads to unnecessary treatment for patients who will not benefit. An example of treatment guidelines that could benefit from more accurate risk estimates is the treatment of cardiac patients with implantable cardiac defibrillators (ICDs). These are expensive devices with invasive implantation procedures that can prevent sudden cardiac death, but most patients prescribed a device do not ultimately use it, and most patients who would benefit from the device are not prescribed one under current guidelines [3]. This poor decision-making is due to the limited discriminative ability of the risk factors these guidelines are based on [14]. ICDs are one of many treatment options throughout medicine that would benefit from more accurate models of patient risk.

Efficient allocation of limited resources in hospitals keeps costs manageable and ensures that patients receive the best care possible. This is important both at the provider level, where limited resources such as personnel and equipment should be used as effectively as possible, and at the level of healthcare systems. Models that can identify high risk patients not only identify which resource allocations will lead to the best outcomes, but allow measurement of where hospitals are over or under-utilizing resources to identify areas for improvement.

Patient outcomes are increasingly used to evaluate healthcare providers, and measures of hospital performance such as 30-day readmission are being tied to hospital reimbursement by Medicare and Medicaid as well as made available to the public. Accurate estimates of a patient's baseline risk of adverse outcomes are essential to fair hospital evaluations, as it is known that outcomes-based penalties disproportionately affect institutions that serve high-risk populations [42, 41, 57]. Penalization of these institutions exacerbates the challenges of serving high-risk patients, and can unfairly impact low income communities. Models that accurately measure a patient's baseline risk of poor outcomes, independent of the quality of care provided, improve the fairness of these policies and provide more actionable information to institutions that allows them to improve their patient outcomes.

The application of state of the art machine learning methods can improve the accuracy of clinical risk models. These methods can make better use of patient data during model development, resulting in models with more accurate predictions. They can also be used to learn novel discriminative risk factors that improve prediction when included alongside other risk factors in a model. The use of better learning methods in the development of clinical risk scores would result in a more effective overall healthcare system.

Yet there are properties of clinical prediction problems that challenge standard learning approaches, and require the development of new methods and innovative applications of existing ones. These properties include highly imbalanced datasets due to rare outcomes, small patient populations in clinical trials, the availability of long-term time series data for prediction, an abundance of missing values, and the need for models that can be interpreted and validated by clinical experts. In this proposal I address several of these challenges and demonstrate improvements to clinical risk stratification through the application and development of machine learning methods. These improvements come from two directions: through the identification of novel risk factors from physiological recordings, and through better methods for training and applying risk models when patients do not have all risk factors available, as is common in real-world clinical data.

1.1 Incorporating Long-Term Physiology into Risk Models

Most standard risk scores make exclusive use of point-in-time measurements for diagnosis and risk stratification [1, 28, 75, 44, 2, 16]. These snapshot measurements are associated with a patient's condition and their risk of outcomes, but do not fully characterize the patient's physiology. Many measurements, like heart rate, blood pressure, or blood glucose, vary not only according to daily cycles, but also in response to a range of stimuli including stress, exercise, and sporadic physiological occurrences like cardiac arrhythmias. Point-in-time measures lose predictive power due to variance caused by the patient's state and the time of day when measured. However they also fail to capture trends over time or abnormal responses to particular conditions or stimuli that provide many more insights into a patient's condition.

Studies have shown that considering measurements of heart rate and blood pressure over time using ambulatory recordings is more useful for identifying high-risk patients than point-in-time measurements [72, 56, 40]. Historically, the inability to collect and analyze long-term recordings from large numbers of patients, and the infeasibility of applying risk scores derived from these recordings in standard clinical practice has led to a paucity of work investigating the use of these recordings for risk stratification. While the limited body of work in this area has identified potential risk markers from these recordings [5, 11, 68], it is an open question what the most useful features are to extract from these data for prediction of adverse outcomes. Optimizing the use of these recordings for risk-stratification is complicated by the difficulty of manually identifying high-risk properties in these recordings. It is infeasible for humans to identify subtle and infrequently occurring indicators of risk that are present in data given the length and quantity of recordings. Although the costs to implement long-term ambulatory monitoring of physiology continue to decrease, making the use of these recordings in standard practice more feasible, this monitoring is still more expensive and time-consuming that point-in-time measurements collected in the clinic. In order to justify the use of this data in standard clinical practice, the demonstration of larger improvements over standard risk stratification approaches is necessary.

Computational methods that automatically identify and measure more discriminative risk factors in these recordings can make physiological recordings a valuable tool for patient risk estimation, and could lead to significant improvements in clinical risk stratification. These methods can also identify structure in this complex data that provides insights that improve clinical understanding of physiology over time, and motivates additional clinical research. Previous work on identifying risk factors from time series has lead to improvements across a broad range of clinical problems, including risk stratification of cardiac outcomes [68], detection of epileptic seizures [65, 52], and prediction of adverse outcomes in the neonatal intensive care unit [63]. This thesis advances the body of work on learning risk factors from physiological recordings through two different approaches.

First, we investigate methods for learning a concise representation of a patient's underlying physiological state from the broad spectrum of phenomena that can be observed in a recording. Physiological responses to a wide range of stimuli manifest as sporadically occurring patterns in long-term recordings, and are reflections of a patient's underlying physiological state. It has been shown that in a number of physiological measurements, including measurements of brain activity and heart rate, that there exist a large number of short patterns, occurring on the order of seconds. These sporadic occurrences have associations with a patient's condition that traditional time series statistics fail to capture. However these short physiological patterns, while informative, can be so plentiful that it is a challenge not only to train models effectively, but also to glean insights into the structure of the data due to the quantity and complex correlation structure of these patterns. In this thesis, we show that by identifying organization and hierarchical structure in these short patterns in recordings, it is possible to uncover a lower-dimensional, multi-faceted representation of a patient's physiological state. This representation of latent state is demonstrated to have utility for more accurate prediction of patient outcomes, while also providing an interpretable and easily visualizable representation of recordings that can provide clinical insights into the relationship between physiology over time and patient outcomes.

As a second direction for improving the learning of risk factors from recordings, we consider the use of more powerful learning approaches for feature extraction from this data to identify more discriminative features. Prior work on learning risk factors from these recordings is based on measuring or combining predefined features on the data. We investigate methods for feature learning that automatically identify collections of features from large amounts of data. Specifically, we consider the application of convolutional neural networks (CNNs), a class of methods responsible for recent major advances in computer vision and speech analysis, to learning features from physiological recordings. CNNs are capable of learning arbitrarily complex representations of the data, and a single model can learn and integrate features at a variety of time scales, from seconds to minutes to hours. We explore the space of CNN methodologies for learning features for clinical risk stratification, and show significant improvements to risk estimates using the derived features.

The approaches studied in this thesis for learning improved risk factors from physiological recordings were investigated in the context of two very different clinical problems.

Cardiovascular disease is the leading cause of death in the United States, affecting 35% of Americans over the age of 20, and is the diagnostic group responsible for the largest percentage of total health expenditures, resulting in an estimated 315 billion dollars in costs in 2012 [31]. It is projected that by the year 2030 the annual cost of cardiovascular disease will rise to over 1 trillion dollars [36]. Cardiac risk stratification is essential for guiding treatment decisions, and can have a large impact on patient outcomes and costs. However, long-term electrocardiographic (ECG) data remains to be utilized fully in the prediction of adverse events. Approaches to leveraging this data are limited, and are centered around specific hypotheses about high-risk patterns in the data. The increasing availability of large quantities of recordings and the limitations of expert knowledge present an opportunity for the learning of novel predictive information from this data. Chapter 2 discusses the learning of higher-level structure from the occurrence of short patterns in 24 hour heart rate recordings. This chapter presents a specific methodology called heart rate topic models, and demonstrates that the high-level structure identified by the models not only provide novel predictive information, but that these models are superior to previously proposed heart rate risk scores for prediction of cardiac death. Chapter 3 presents an application of convolutional neural networks to learning risk factors from the same heart rate recordings. The networks learned achieve significantly better prediction of cardiac death from heart rate recordings than previously proposed heart rate measures, including the topic models presented in chapter 2. The features learned by the CNN were able to improve overall prediction of cardiac death even when considered in conjunction with standard risk scores used in clinical practice, and were more discriminative than standard clinical risk factors.

Analysis of sleep recordings is a critical tool in the diagnosis, treatment, and understanding of sleep disorders, which have been estimated to affect over 50 million Americans [38]. Sleep staging, the classification of windows of sleep EEG (electroencephalogram) data to evaluate a patient's quality of sleep, has traditionally been done manually. Automated methods for sleep staging generally try to mimic human classification of the EEG, however the standard definition of sleep stages is extremely simple and does not account for variations in EEG properties across patients. Chapter 2 describes the use of topic models to learn latent state from short patterns in sleep EEG in an entirely data-driven and patient-specific way. These models provide a more nuanced representation that identifies structure in sleep missed by the current sleep staging standard, and offer a complementary approach to evaluating sleep data.

1.2 Addressing the Real-World Prevalence of Missing Values

Commonly used clinical risk scores are not designed for use with missing covariates. This leads to two major limitations. First, when estimating the risk for a patient, these models cannot be used as intended if one or more model covariates are unknown for a patient. Risk models that can generate accurate estimates even when not all useful variables are available would increase the value of model predictions. Second, these multivariate risk scores are developed to include only risk factors commonly collected for all patients. However additional useful measurements are often available for a subset of patients, which may also be collected depending on the patient's condition, the judgment of their physician, or the availability of resources at their healthcare provider. The ability for risk models to incorporate these additional risk factors when available would improve the accuracy of their risk estimates. Unfortunately, the most common method for learning clinical risk models, logistic regression, is not designed for the setting where model covariates are missing either during training or when evaluating patients. This property is true for most standard algorithms for classification and regression, and methods that can readily handle missing values (e.g. random forests, k-nearest neighbors) lack interpretable models, a critical limitation in the medical domain where understanding model parameters and expert validation of models is essential.

This limitation is troubling given the prevalence of missing data in real-world clinical settings: sensors frequently malfunction, equipment may be unavailable, the severity of a patient's condition can prevent measurement of variables, or measurements may not be recorded by hospital staff. Ignoring records with missing values when developing a model can lead to serious biases that limit their effectiveness on real patient data [69], and ignoring records with missing values at test time makes risk models unusable for large percentages of patients.

Missing values in data when analyzing clinical trial data or training predictive models are generally handled by imputation, the filling in of absent values with estimates. A multitude of papers have been published on identifying the best imputation method to use in different clinical settings, however nearly all of these focus on interpreting the results of clinical trials. Less work has investigated the effect of missing value handling during the development of predictive models, and these works treat the imputation and risk model learning as two separate processes. Several studies have found that considering these steps separately hurts the quality of models and the accuracy of predictions, and that handling the imputation and risk model learning in conjunction leads to improved prediction [47, 22].

Additionally, approaches to handling missing values in the analysis of clinical data assume that the data are not missing at random, even though this assumption often does not hold [67]. Complex processes govern whether various measurements are collected, depending on the patient's condition, the judgment of individual clinicians, and resource availability at institutions, and in many cases the variables with missing values have structured missingness that cannot be explained by the observed variables. Incorrectly assuming data are missing at random not only leads to inaccurate estimates of missing values that lead to inaccurate estimates of risk, but also introduces noise into the training of the risk model that can hurt discrimination even on records with all variables observed.

This thesis works towards addressing these limitations by developing a general purpose method for jointly learning imputation models and classifiers. This method considers the effect of the imputation quality on classification loss and vice versa, to improve the accuracy of risk models in the presence of missing covariates. Unlike existing methods for jointly learning imputation and classification models from the machine learning literature, this method is designed for use with data of many real-world medical settings, which is often not missing at random, and can be used with a variety of regression or classification models.

We evaluate the utility of the proposed method in the context of identifying high-risk trauma patients at the time of admission to the hospital. Trauma is the leading source of death for Americans under 44 years of age [26]. Trauma patients are some of the highest risk patients admitted to hospitals, and the urgency of their care leads to an abundance of missing clinical observations at admission to the hospital. Chapter 4 evaluates the proposed method on a large national registry of trauma patients, and finds that it significantly improves prediction of several adverse outcomes relative to standard imputation approaches.

1.3 Thesis Organization

The thesis is organized as follows.

- Chapter 2 presents an application of topic models for learning higher-level structure in physiological recordings, and investigates its use for prediction of cardiovascular death from heart rate recordings, and for the unsupervised identification of sleep structure in electroencephalographic recordings.
- Chapter 3 presents the use of convolutional neural networks to learn features for the prediction of cardiovascular death from heart rate recordings.
- Chapter 4 presents an approach to jointly learning imputation and classification models for use with not missing at random data, and evaluates the approach for the prediction of adverse outcomes in trauma patients in a large national registry.

CHAPTER II

Learning Latent States in Physiological Recordings

2.1 Physiological Topic Models

2.1.1 Introduction

Recordings of physiological data contain a wealth of information about a patient's health and their risk of adverse outcomes. They reflect the workings of the body over time, in a way that point-in-time measurements used in clinical decision making are unable to capture. Long-term recordings of physiological data spanning hours or days can capture variations and trends in physiology over time that may not be evident over shorter durations, and can measure physiological responses in a natural setting to stress, exercise, and other stimuli. These varied observations provide glimpses into a patient's underlying physiological state.

Many kinds of predictive structure have been identified in long-term physiological recordings, although most work has focused on standard time series analytic methods at time scales on the order of minutes to hours. It is known that in several kinds of physiological recordings there exist short patterns, on the order of seconds, that carry important information about a patients underlying physiological state [33, 68]. Risk models based on the frequencies of these patterns have been shown to be predictive of adverse outcomes and to add complementary information to measures that identify structure at larger time scales [64, 68]).

It is possible to identify arbitrarily many patterns, each weakly associated with patient risk, and with many correlations among them. These patterns can be prohibitively large in number, occur infrequently, and may carry different implications depending on the other patterns they occur with. The large number of weak features makes models based on these pattern frequencies difficult to understand and interpret, and they are challenging for clinicians to visualize across a recording.

These patterns provide different views of a patient's underlying physiological state. Computational models to analyze and summarize the many observable patterns in these recordings can uncover a multi-factorial representation of a patient's latent physiological state that goes beyond simply counting frequencies of individual features. The representation learned captures predictive structure in the cooccurrence of these patterns that can be used to improve estimates of patient risk.

Computational methods also have a role in making long-term recordings more interpretable to clinicians, as manual analysis is time-consuming and challenging. A representation of a patient's latent state can be used to create visualizations that facilitate understanding of these recordings both for individual patients and across populations. Identifying unusual characteristics of long-term recordings can be like finding needles in a haystack, and even when long-term recordings become part of standard clinical practice manual analysis of the recordings is not feasible. These models can summarize large amounts of data into concise representations that simplify analysis and interpretation of recordings and risk factors by clinicians.

The work in this chapter addresses these challenges with an application of topic models to the analysis of physiological recordings. Topic models, a class of methods from natural language processing, summarize collections of documents by identifying semantically related sets of words called topics and associating each document with a mixture over these topics. The topics learned, and the ability to associate a document with a mixture over topics, allow for additional kinds of analysis like clustering, indexing, or summarization. While topic models have been used in a variety of application domains, in this work we extend the use of topic models to physiological time series data. This application is motivated by the presence of many physiologically meaningful short patterns in the data, which we consider as analogous to words in a document, from which meaningful higher-level structure (topics) can be learned when aggregated over a long recording.

The representation of physiological recordings as mixtures of topics has several advantages. Topics give a concise description of long recordings, providing a useful visualization of large amounts of physiological data by summarizing patients with a distribution over topics. The approach finds sets of patterns that tend to cooccur with one another across patients, providing insights into the relationships between different patterns and their organization in the recordings. It can distinguish between different occurrences of a pattern based on their context (i.e., the other patterns that occur with it), giving a more sophisticated understanding than simply counting patterns. Topic models complement time-sensitive approaches like hidden Markov models, which have difficulty associating patterns that occur far apart in time and assume discrete states at each time point, through the "bag of words" assumption that relates all pattern occurrences in a document.

Several previous works have investigated the discovery of states in physiological time series to improve analysis of physiological recordings. Cohen et al. modeled physiological states in multivariate intensive care unit (ICU) time series using a hierarchical clustering approach [19]. Saria et al. used a hierachical graphical model to simultaneously model multivariate time series and learn a set of disease "topics" in neonatal intensive care unit time series, using mixtures of vector autoregressive processes [62]. In contrast, the proposed approach does not constrain the choice of data representation. The patterns considered by the model can be univariate or multivariate, could span a range of time scales, or could be based on time or frequency domain features. This flexibility makes the approach applicable to a broader range of problems.

2.1.2 Latent Dirichlet Allocation

While many approaches to topic modeling have been proposed, latent Dirichlet allocation (LDA) is by far the most widely used. LDA is a generative model that assumes that a collection of documents was derived from a set of topics (e.g., the proceedings of an artificial intelligence conference may consist of topics such as planning, knowledge representation, or machine learning) [9]. Each topic underlying the corpus can be characterized by a distribution over words in the vocabulary: a topic about machine learning may have high probabilities of words like “unsupervised” or “classifier”. Given only a set of documents and their respective word frequencies, the LDA generative model can be used to learn a set of topics to explain the corpus, and attribute each document to some mixture of these topics.

More concretely, LDA defines an underlying set of K topics, where each topic k can be defined by a distribution over all of the words in the vocabulary. Each document d is itself generated by a distribution over topics θ_d, where the generative process assumes two steps for each word w_{di} in the document. First, a topic z_{di} is chosen by sampling a topic from that document's topic distribution θ_d. Then the word w_{di} is sampled from that topic's distribution over words β_k. More concretely, the model defines for a document d the prior distributions over observed words w_{di}, their latent topics z_{di}, and the document distributions over topics θ_d as:

$$\theta_d \sim Dir(\alpha)$$

$$z_{di} \sim Multi(\theta_d)$$

$$w_{di} \sim Multi(\beta_{z_{di}})$$

Where α parameterizes a symmetric Dirichlet prior on topic distributions. In training the model, the parameters of interest are the θ_d and β_k values, which characterize the topics of semantically related words and the proportions of these topics with respect to each document.

2.1.3 Applications of Physiological Topic Modeling

We study the application of topic models to physiological recordings in two different clinical problems.

In Chapter 2.1, we consider the use of topic models to identify a novel risk factor. We present an approach for the identification of topics in heart rate recordings, and evaluate the approach for prediction of cardiovascular death (CVD) following an acute coronary syndrome. We find that this approach improves prediction of CVD relative to previously proposed heart rate risk factors. Much of the work in this section has been published previously [23, 24].

In Chapter 2.2, we consider topic models as a way to learn a data-driven, patient-specific representation of sleep recordings. We find through an evaluation on a publicly available sleep EEG dataset that the topics learned not only capture the structure present in the standard sleep stage system, but provided a more informative and nuanced picture of a patient's night of sleep. The work in this section has been published previously in [25].

2.2 Heart Rate Topic Models

2.2.1 Risk Stratification for Cardiovascular Death

Cardiovascular disease affects 7% of adults in the United States and is the single leading cause of mortality [60]. It is estimated that in 2011, there were over 1 million coronary attacks in the U.S., with nearly one-third of individuals suffering a coronary attack dying of it. A similar situation exists in other parts of the world, in particular developing countries, where it is estimated that by the year 2020 cardiovascular disease will be responsible for 40% of all deaths [59]. As a result, a wide variety of treatments (e.g., procedures, medications, devices) have been developed to prevent future events and reduce an individual's risk of mortality.

Despite the wide array of these treatment options, determining the appropriate level of therapy for an individual remains challenging. Underestimating a patient's risk for cardiovascular death (CVD) can preclude the use of potentially life-saving preventative treatment. Conversely, overestimation of risk can lead to application of expensive and often risky procedures which provide little or no benefit to the patient. One example of this is the implantable cardiac defibrillator (ICD), an expensive device requiring surgical implantation capable of preventing fatal arrhythmias. The outstanding majority of the patients receiving an ICD under current guidelines do not benefit from it, while most patients who die of arrhythmias are not prescribed one [3]. Accurate risk stratification is vital for matching patients to appropriate treatments, with the potential to both save lives and reduce health care costs.

Risk stratification is especially important following an acute coronary syndrome (ACS). ACS refers to range of conditions caused by insufficient blood flow to the muscle of the heart. The resulting lack of oxygen can result in temporary dysfunction of the heart, and in more severe cases death of muscle tissue (referred to as a myocardial infarction). Damaged heart tissue impairs the conduction of the electrical impulses that control the functioning of the heart, increasing the risk of cardiac arrhythmias, some of which can be life threatening. The risk of adverse cardiac outcomes is higher following an ACS than in the general population, however an individual patient's risk varies widely due to the heterogeneous nature of ACS [10]. Determining the appropriate treatment for a patient post-ACS, which usually consists of some combination of medication, coronary revascularization, and an implantable device, depends on accurate estimates of their risk of various cardiac outcomes.

A number of different biomarkers and clinical scores have been generated to quantify a patient's cardiac risk, however most of these capture information only at a single time point, missing prognostic information in variability and progression in risk over time. These include biochemical markers such as brain natriuretic protein and C-reactive protein; measures derived from echo- and electrocardiography; and clinical risk scores such as the Thrombolysis in Myocardial Infarction (TIMI) risk score, which integrate information related to different patient characteristics. While these metrics can identify many high-risk patients, they utilize instantaneous data and lose potentially valuable information about the performance of the heart over time.

Recent work has focused on addressing this discrepancy and improving patient assessment through novel cardiac biomarkers derived from long-term physiological signals such as the electrocardiogram (ECG). The ECG measures the electrical potential of the heart on the surface of the chest. This simple measure of electrical activity contains a vast amount of information related to the timing of cardiac activity as well as the structure and functioning of the components of the heart and its related systems. This information, combined with the ease with which ECG can be recorded and its prevalence in health care settings, has made it a common data source for development of cardiac risk stratification metrics. With the recent development of better storage technology and practices, the ability to apply more sophisticated analysis to long-term ECG recordings has become possible.

Of particular interest is studying changes in the heart rate over long periods of time, as a way of characterizing abnormal autonomic nervous regulation of the heart that predisposes patients to fatal arrhythmias. Much of the initial work in this space (e.g., heart rate variability [51], heart rate turbulence [4], and deceleration capacity

[6]) relates either aggregate changes in heart rate or the presence of a specific heart rate pattern to an increase in cardiovascular risk.

Most recently, this work has been generalized to identify and integrate information in multiple heart rate motifs (i.e., short heart rate sequences discovered in a data-driven manner from long-term ECG) that are over- or underrepresented in patients experiencing cardiovascular outcomes [18]. While these "heart rate motifs" have been found to be predictive of several cardiac outcomes [68], the number of motifs present in the data is overwhelming, necessitating aggressive feature selection to prevent overfitting that may lose valuable information in the many remaining motifs and their interactions. A representation consisting of tens of thousands of features creates challenges for interpreting the model and identifying trends and structure in the data.

The use of topic models to learn structure in the cooccurrence and relationships between large numbers of motifs may provide additional information about the functioning of the heart relevant to the prediction of cardiac outcomes. A higher-level representation of the occurrences of thousands of motifs over long recordings can simplify learning, uncover additional predictive information, and allow for greater interpretability of heart rate motifs. This chapter builds on these hypotheses and addresses the shortcomings of the heart rate motifs approach by identifying generative structure in the full set of motifs throughout the population. We learned topics consisting of related heart rate motifs capable of characterizing longitudinal ECG recordings across a population. These data-derived topics provide an interpretable abstraction of the heart rate motifs in a recording, while providing a concise, task-independent representation that can be used to better assess overall cardiac health.

The goals of this section are to:

- present an approach for accurate and interpretable cardiovascular risk stratification by using topic models to identify higher-level structure in short-term heart rate structure across a population

- evaluate the performance of these topics in assessing cardiovascular risk alongside a popular risk measure on a real-world dataset consisting of long-term ECG recordings with nearly year-long patient follow-ups.

2.2.2 From Heart Rate Motifs to Heart Rate Topics

Most cardiac biomarkers derived from the ECG focus on the idea that an inability for the heart to adjust its rate to compensate for different physiological situations is indicative of a high risk for cardiovascular problems. In recent work on heart rate motifs, Chia and Syed explored identifying and integrating information from the frequencies with which different short-term heart rate patterns occur in long-term ECG to assess cardiac patients. The approach converted 24 hour heart rate time series measured from long-term ECG data into symbolic sequences, corresponding to low through high heart rate ranges, and discovered short approximate symbolic motifs associated with high or low risk individuals [18]. This approach identified a small set of over- or under-represented symbolic motifs in patients suffering cardiovascular death, and integrated the frequencies of these motifs over an ECG recording into a predictive model. This approach found that by studying the short-term heart rate patterns over long-term data it was possible to identify information complementary to other commonly used clinical variables.

While the heart rate motifs method improves upon existing methods for risk stratification, there are several challenges with the use of heart rate motifs. For even a moderate choice of motif lengths, the number of motifs in the data grows unreasonably (with 4 symbols and motifs of length 8 there are over 60,000 motifs). The use of real-world long-term ECG datasets, which contain limited numbers of patients, necessitates the use of only a sparse subset of features to prevent overfitting. This has two effects: first, this prevents the leveraging of large numbers of patterns and their relationships for prediction. Second, this model describes only a small percentage of the original recordings, and in some cases an individual may have none of the predictive motifs present in their ECG. This provides a limited representation of the data, and can be challenging to interpret. Another difficulty is that while the resulting model has good predictive power, it provides limited insight into the generative structure of these motifs across a recording or a population. The use of individual motif frequencies identifies no structure in the relationships between motifs, and cannot distinguish cases where one heart rate motif can appear in multiple contexts. Understanding the higher-level structure of these motifs may provide novel insights into the data and correspondingly into cardiac physiology. Additionally, because many cardiac outcomes exist (e.g. ventricular arrhythmias, atrial arrhythmias, myocardial infarction, recurrent ischemia), the use of motifs requires selection of different sets of predictors for each one. This makes it challenging to compare predictors across types of cardiac events, because there is no easy way to reconcile the two models.

By extracting underlying generative structure independent of any particular endpoint, it may be possible to retain useful information present in the full set of motifs relevant to overall cardiac frailty, without tailoring the representation to a particular task. We investigate one approach to uncovering latent structure, topic modeling, to heart rate motifs. Topic models have been used in a variety of application domains to identify sets of semantically related words capable of describing a set of texts. In this section, we extend the use of topic models to heart rate data. In particular, we investigate advancing the work of Chia and Syed by first symbolizing and identifying "words" from a physiological time series, and then identifying latent topics in the data. Instead of independently comparing each motif frequency between the ECG recordings for two outcomes, we consider the motifs as words, constituents of a single document corresponding to a recording. After defining a document as the frequencies of all motifs occurring in a recording, we can use this text-like representation to extract higher-level structure in the documents by learning topics over the motifs.

The application of topic models to heart rate motifs has the ability to address the challenges mentioned. By modeling the heart rate motifs in a recording with a set of generative topics, every motif in the data is explained by a contribution from some topic, accounting for the entirety of every patient's recording. By relating symbolically different but functionally similar motifs, all patterns in the data can be leveraged to identify useful predictive information. This allows the previously unwieldy motif frequency representation to be condensed into a mixture over a small set of topics, providing a similarly sparse representation but with the potential for greater predictive power.

This representation of the data as a mixture of topics also serves to identify generative structure behind the motifs. It allows the analysis of related sets of patterns, and easier identification of trends across recordings or patients by assessing the variability among a few topics rather than a large number of motif frequencies. The unsupervised nature of topic models means that this condensed representation provides information about the full set of motifs without focusing on a particular outcome. After identifying latent heart rate topics, the patient mixtures over topics can be used across a variety of cardiac events, allowing for a quantification of the relationships between their differing predictors. For these reasons, we believe that topic models provide a better interpretation and understanding of heart rate motifs, while maintaining and potentially expanding upon the predictive power of these features for cardiac assessment.

2.2.3 Methods

The proposed method consists of four stages. First, a heart rate time series is generated from the ECG signal. This involves identifying QRS complexes and then calculating time between adjacent complexes. This is done using the open source QRS detection algorithms proposed by Hamilton et al. [34] and Zong et al. [76] to identify

QRS complexes at time instants where both algorithms agree. The instantaneous heart rate was defined as the time between all pairs of normal sinus beats.

Second, this time series is converted into a symbolic sequence by dividing each patient's heart rate values in equiprobable bins (e.g. quartiles), using the SAX method [48]. We apply SAX to each patient's time series separately. This provides robustness to inter-patient variability in baseline heart rates in that a symbol corresponding to high heart rate has a consistent meaning across individuals, an essential property when learning structure across a population.

Third, we count all occurring motifs of a given length (a motif of length four might be: low, high, low, high), each of which describes a heart rate pattern. We consider all substrings of a given length in this symbolic sequence as words within the document representation, characterizing long-term ECG as a bag of motifs and their respective frequencies of occurrence. For a given motif length n, we calculate the frequency of all n-length substrings in the symbolic sequence, allowing for overlaps. In earlier work on heart rate motifs, a formulation of approximate motifs was used, in order to group functionally equivalent heart rate sequences as well as to account for the hard symbolization boundaries. This definition of motifs required an extensive framework to efficiently compute motif frequencies on large datasets. The application of topic models, which provide a natural and robust way to identify groups of related motifs, allows us to sidestep the challenges involved in computing approximate motif frequencies.

Finally, we use the motif frequencies for each patient across a population to train a topic model using LDA than can be used for cardiovascular risk stratification. We consider each patient's ECG recording as analogous to a document, with each heart rate motif corresponding to a word. It is then possible to treat the set of symbolized heart rate time series across the patient population as a corpus, over which we can identify topics of heart rate motifs. The LDA parameters were estimated through variational Bayesian inference as described in [9]. We used open source software for variational inference-based estimation of the model parameters from the symbolized heart rate corpus, resulting in a set of heart rate topics and each patient's topic distribution.

2.2.4 Evaluation

We evaluated the clinical utility of heart rate topic models for predicting cardiovascular death (CVD) using data from patients admitted to hospitals with non-ST-elevation acute coronary syndrome from the TIMI-MERLIN36 trial (for more details on this data, see Appendix A). For each patient, the first 24 hours of continuous ECG were used to extract the heart rate time series.

The topic model estimation was performed on the full set of patient recordings. We experimented with varying values of motif length (n=2,4,6,8), alphabet size (A=2,4,6,8), and number of topics (K=1,. . .,10), accounting for variability in model estimation by repeating the LDA parameter estimation process 10 times. The choice of n and A was made using data from DISPERSE2 as a validation set, while selection of K and choice of model over parameter estimation replicates was done using the Bayesian information criterion (BIC). This corresponded to a motif length of 6, an alphabet size of 4, and 10 topics, which was used for the full set of experiments.

A CVD risk score was developed by training a logistic regression model using each recordings mixture over topics as features. The training was done using leave one out cross-validation. The same leave one out cross-validation approach was used to develop a CVD risk score using heart rate motifs. A motif length of 6 and an alphabet size of 4 were used for the motifs to provide a direct comparison to the topic models. The regularization parameters for the logistic regression models were selected using data from DISPERSE2.

We evaluated whether the model using topic mixtures associated with each patient's recording improved prediction of CVD. First, we compared to the method of heart rate motifs, to assess whether the additional structure derived by the topic models provides comparable or additional information to the raw motif frequencies. This evaluation was done on the full set of patients with heart rate recordings available (n=5136). Second, we compared the topic-based score to a range of cardiac risk factors, including previously proposed heart rate measures (deceleration capacity [6], heart rate turbulence [4], and heart rate variability [51]), as well as against standard clinical risk scores. This included the TIMI risk score (TRS) [1], which is commonly used to evaluate a patient's risk of adverse cardiac events, and incorporates an array of clinical risk factors (e.g., age, weight, blood pressure), serum cardiac biomarker levels, number of anginal events in the previous 24 hours, and several other factors. TRS provides an easily calculated score that a clinician can use to assess a patient's cardiac risk. Due to the ease of use and prevalence of the metric, the TRS can be considered a standard for risk stratification. We also compared against left ventricular ejection fraction (LVEF), an echocardiographic measure that is one of the most important factors in clinical decision making following an ACS. Evaluation with these additional scores indicates whether the topic model contains predictive information that is not captured by standard heart rate or clinical measures. Due to additional signal quality constraints for some heart rate measures, and missing measurements for LVEF, this evaluation was conducted on a smaller subset of patients (n=2599).

Risk scores were compared in terms of their ability to predict CVD using the area under the ROC curve (AUC). Additionally, Cox proportional hazards models were used to measure the ability of these scores to predict CVD when considering the timing events and censoring of outcomes in the MERLIN trial due to the limited follow-up duration. To achieve a dichotomized predictor (a binary value corresponding to low or high risk) for use with the proportional hazards model, the predictions from the logistic regression model were dichotomized at the top quartile, to be consistent with prior studies. We also evaluated the benefits of adding the topic score to the set of existing risk measures when combined into a single risk score using logistic regression. This improvement was measured in terms of AUC, integrated discrimination improvement (IDI), net reclassification improvement (NRI), and hazard ratios. Calibration was assessed using the Hosmer-Lemeshow χ^2 score. A backwards stepwise elimination process was also used to identify which variables were most useful for prediction, as many of the heart rate measures had non-negligible correlations with one another. The threshold for variable inclusion in the model was set to $p < 0.05$. For a detailed explanation of the evaluation metrics used in this section, see Appendix B.

In addition to quantifying the ability of the topic representation to predict CVD, we investigated the structure learned by the model by inspecting the topics identified and visualizing their distribution across patients.

2.2.4.1 Topic Analysis

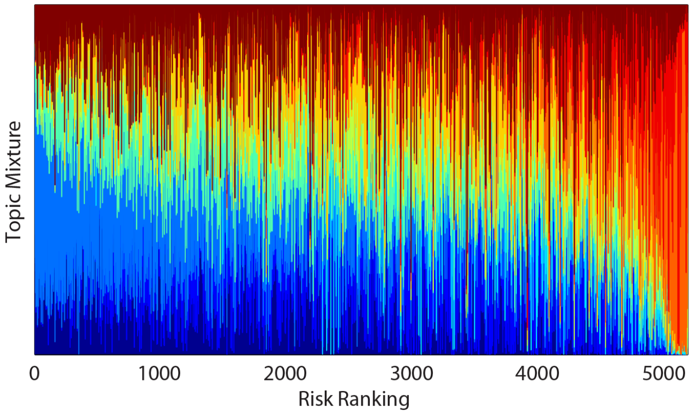

Figure 2.1: Topic mixtures for the testing set of patients. Each patient's mixture of topics is represented by a vertical strip, with patients in increasing order of risk as estimated by the CVD prediction model.

The topic distributions for individual patients varied significantly across the population, as shown in Figure 2.1. This indicates that despite a number of similarities in the symbolized heart rate time series between patients (e.g., many of the patients had long runs of the same symbols, such as “1111”), the model still managed to identify topics that vary significantly across the patient population. As Figure 2.1 shows, the patient mixtures over topics characterize their day-long recordings in a way that is easy to visualize and compare across individuals and subgroups. The ability to reduce each patient into a small number of topics that can be visualized provides a representation of higher-level inferred state that is more compact and easier to understand (i.e., in terms of what topics differentiate patients) than simultaneously displaying the frequencies of a large number of motifs.

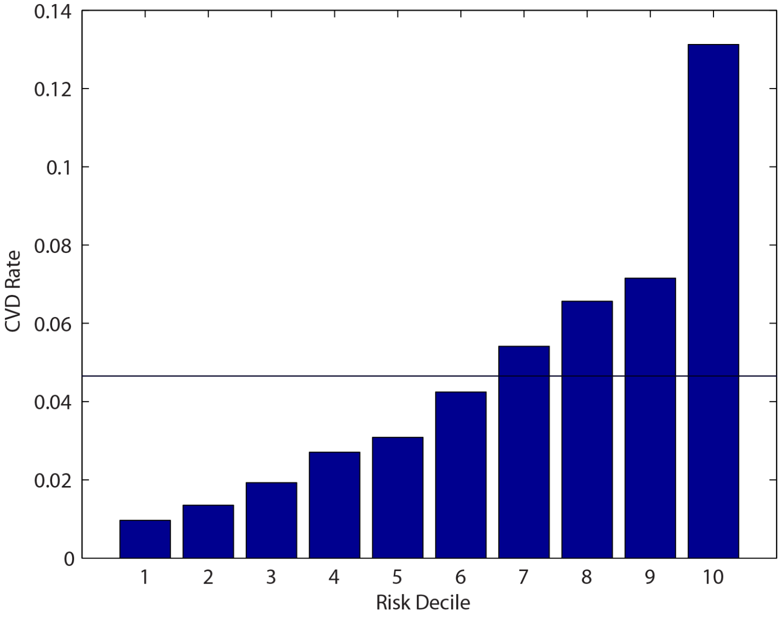

Figure 2.2: Risk of CVD calculated over 10 bins corresponding to estimated risk deciles and the overall population risk (black line).

Figure 2.2 shows the rates of events in risk deciles of the topic score. There was substantial variation in the risk across groups, and the score identified several groups with substantially lower or higher rates of CVD than the overall rate in the cohort.

In a traditional application of topic models, understanding the meaning of a topic is done by investigating its most probable words. For example, a topic about cardi-

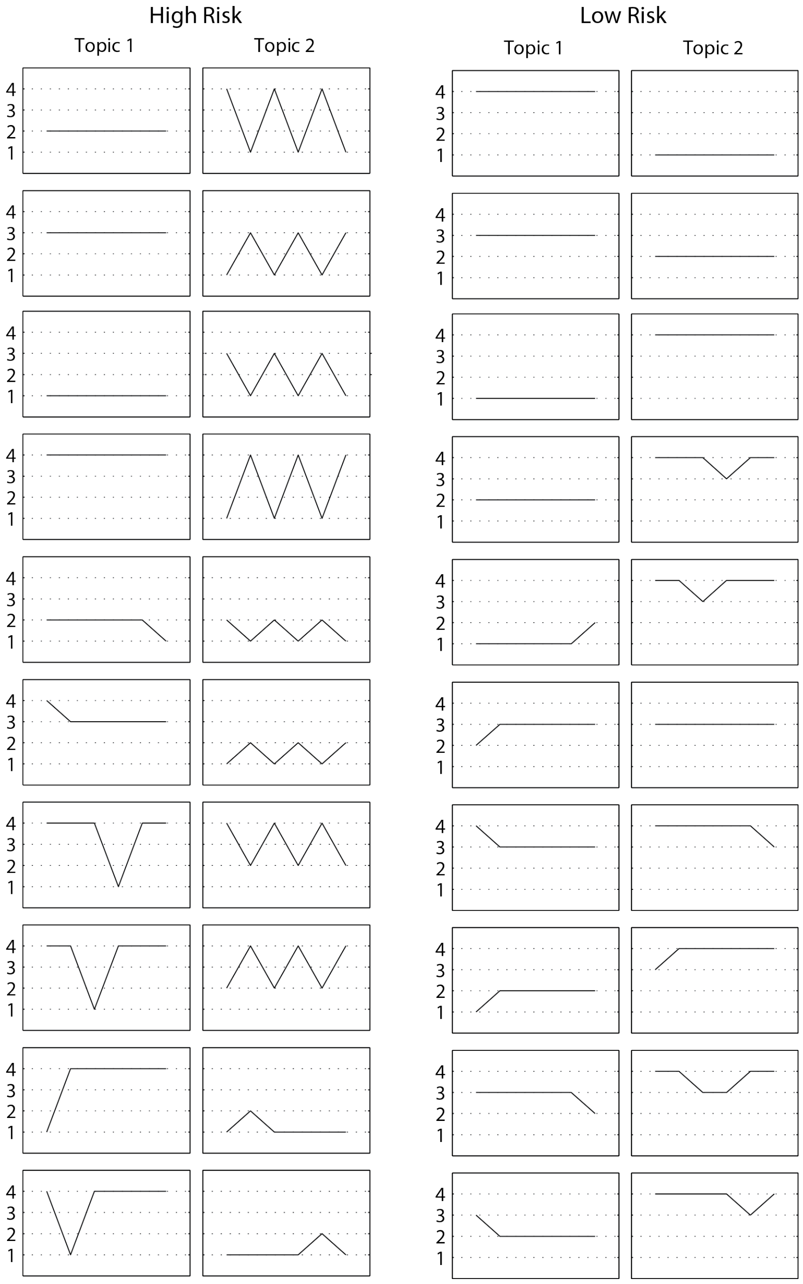

Figure 2.3: Ten most likely motifs for two of the topics most associated with high risk (left two columns) and low risk (right two columns).

ology may have "heart", "ECG", or "myocardial" as some of its most likely words. To gain some understanding of the higher-level structure the model captures, we investigated the most probable words in several of the topics. Figure 2.3 depicts the 10 most likely words in four topics, two corresponding to high risk of CVD (left)

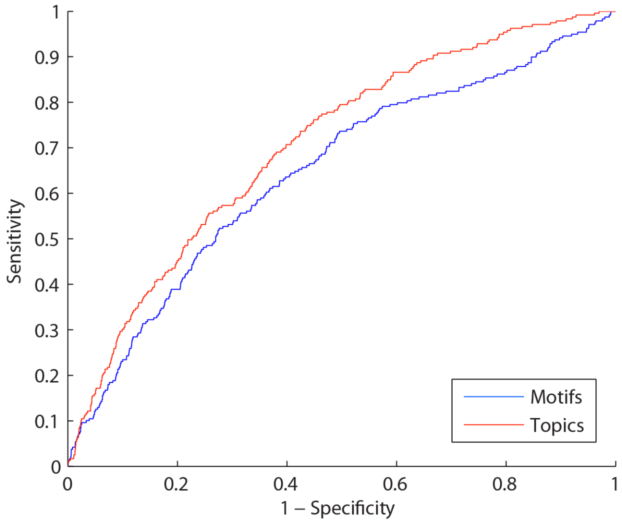

Figure 2.4: ROC curves for the motifs score and the topics score on the full set of patients with heart rate available (n=5136).

and two corresponding to low risk (right) under a Cox proportional hazards model. Topic 1 includes a number of words consisting of a high heart rate (denoted by 4s) interrupted by a single low rate interval. Topic 2 consists of oscillations between heart rates, particularly sharp changes between low and medium-high to high rates. In contrast, the low-risk topics show more stability, with deviations from the baseline having a low amplitude. Long runs of the same symbols occur commonly in several of the topics, showing the ability of the model to differentiate between instances of the same motif depending on their surrounding patterns. Rather than identifying only whether or not a motif is used, the topics model approach uses context to select the most appropriate topic for a given motif instance.

2.2.4.2 Comparison of Heart Rate Motifs and Topics for CVD Prediction

We compared the score learned using heart rate motif frequencies with the score learned using topic mixtures on the full set of patients with heart rate recordings (n=5136). Figure 2.4 shows ROC curves for the two methods. The ROC curve for

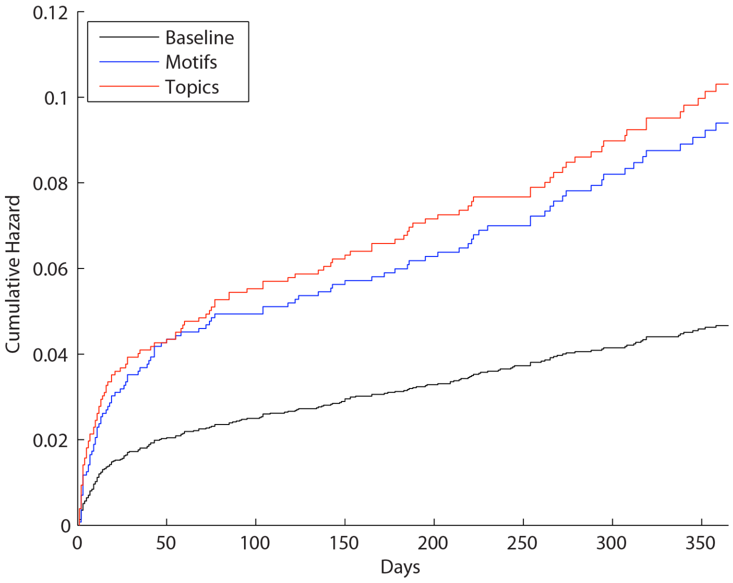

Figure 2.5: Cumulative hazard functions over the first year for the overall population and patients in the highest risk quartile for the motifs and topics scores respectively (n=5136).

the topic score dominated the curve for the motif score. The topic score also achieved a significantly higher AUC than the motif score on this set of patients (Topic AUC: 0.706, Motif AUC: 0.647, p = 0.010). The cumulative hazard functions across the first year of follow-up are shown in Figure 2.5 for the motif score, the topic score, and for the overall population. The cumulative hazard functions for both models high risk groups were much higher than the hazard for the overall population. The topic score had a higher cumulative hazard across the majority of the year. These results indicate that the topic models not only provided a more concise and interpretable representation of the motifs, but that the structure identified provided additional prognostic information.

2.2.4.3 Including Heart Rate Topics with Cardiac Risk Factors

Table 2.1 shows for each risk score evaluated the AUCs, hazard ratios, and hazard ratios adjusted for the clinical scores, when evaluated on the subset of patients with

	AUC	Hazard Ratio	Adjusted Hazard Ratio
TRS	.672	2.7	n/a
LVEF	.708	5.3	n/a
DC	.667	2.9	2.1
HRT1	.619	1.8	1.6
HRT2	.646	2.2	1.4*
HRV	.669	2.8	2.1
Motifs	.659	2.4	1.9
Topics	.685	2.9	2.2

Table 2.1: AUCs, hazard ratios, and hazard ratios adjusted for the clinical scores TRS and LVEF, displayed for all clinical and heart rate measures evaluated. All hazard ratios were significant at the $p < .05$ level except for those denoted by an asterisk.

	Heart Rate Model	with Topic Score Added
AUC	.700 (Ref.)	.709 (.390)
IDI	Ref.	.003 (.034)
NRI	Ref.	.037 (.049)
Category-free NRI	Ref.	.153 (.056)
HLχ^2	3.5 (.899)	6.0 (.650)
Hazard Ratio	2.84 (<.001)	3.25 (<.001)

Table 2.2: Comparison of model containing all previously proposed heart rate measures with a model that additionally contained the topic risk score.

all measures recorded (n=2599). The topic score had a higher AUC than all other heart rate measures evaluated (AUC=.685), and was tied for the highest hazard ratio (2.9). The significance of the topic scores hazard ratio remained even after adjusting for the clinical risk scores. These results suggest that the topic score is not only more predictive than the raw motif frequencies, but achieved better discrimination than any of the investigated heart rate measures. The AUC for the topic score was even higher than for TRS (AUC=.685 vs. .672).

Table 2.2 shows various evaluation metrics comparing models containing all previously proposed heart rate measures with and without the inclusion of the topics score.

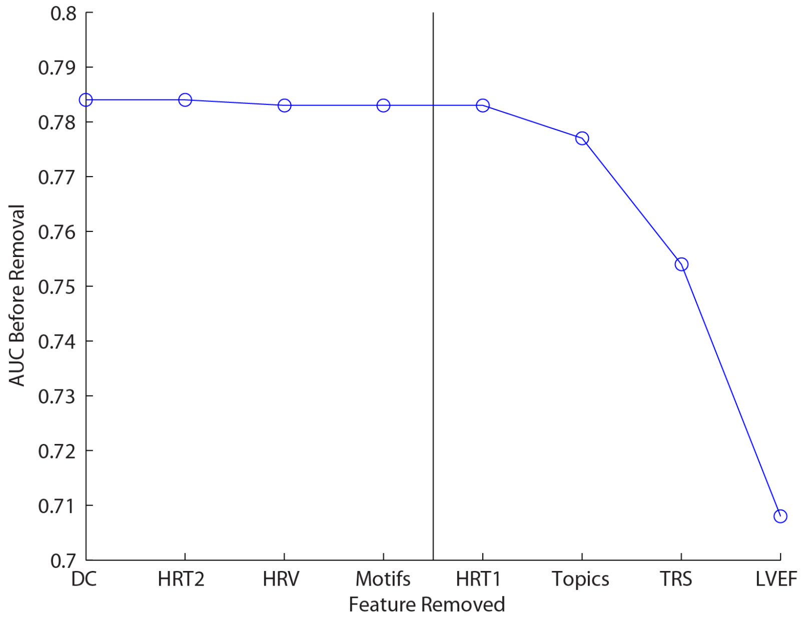

Figure 2.6: AUCs for backwards stepwise elimination over the full set of risk scores (n=5136). Vertical black line marks where the p value threshold of 0.05 was reached.

When added to a risk model containing all previously proposed heart rate measures, the topic score improved the model AUC slightly, but not significantly (AUC = 0.70 vs. 0.71). However the inclusion of the topic score resulted in significant improvements to the IDI and NRI (IDI: 0.003, p=0.034; NRI: 0.037, p=0.049), as well as an improvement to the category-free NRI (NRI=0.153, p=0.056). The hazard ratio also increased (Hazard Ratio=2.84 vs. 3.25). Both models were well-calibrated (HL-$\chi^2 > 0.05$). These results suggest that the topics score not only achieves better discrimination than the other measures individually, but identifies additional information about patient risk that the full set of prior measures did not capture.

Figure 2.6 shows the AUCs and variables removed at each step of the backwards elimination process when all risk scores were considered as variables in a logistic regression model. When included with the full set of risk scores, the topics score was retained in the final model (p<0.001). It was one of only two heart rate measures selected (along with heart rate turbulence onset, p=0.024), suggesting that the topics score is a preferable choice to the other heart rate measures when building a comprehensive risk model including clinical information. The clinical scores were the most significant variables in the model.

2.2.5 Discussion

The use of topic models on long-term heart recordings leverages large numbers of different patterns in the heart rate to uncover higher-level structure in the heart rate. The topics, which were learned without using information about patient outcomes, resulted in a small, task-independent representation of the recordings. The topics were not only more accurate at CVD prediction than heart rate motifs, but achieved higher discrimination than all other heart rate measures investigated. When combined with the full set of heart rate measures, topics improved prediction, suggesting that the improved discrimination is at least partly due to the ability of the topic models to identify novel structure in the recordings associated with risk of CVD. When considered in a backwards stepwise elimination model alongside standard clinical risk scores the topics score was retained in the final model and was the most significant heart rate metric, indicating that its utility remains in the presence of clinical measures and it was preferable to other heart rate measures. Importantly, this improvement in performance is accompanied by an increase in the ability to visualize and interpret cardiac activity with topic models.

There are several limitations to this work. First, we trained the topic models without supervision to allow for a task-independent representation. Integration of patient labels into the topic learning process, using a method such as the one described in [8] could improve upon the discriminative ability of the topics and give a more favorable comparison with alternative approaches. Second, further investigation of the methods on a larger dataset is needed to validate the approach's performance. Using data with longer follow-ups would also provide a more accurate analysis of the risk stratification performance. Finally, attributing physiological significance to the heart rate topics could provide more insight into what aspect of the topic representation is leading to an improvement in prediction, and could inspire further work.

2.3 Sleep Topic Models

2.3.1 Sleep Analysis

Sleep disorders, which include conditions such as sleep apnea and insomnia, have been estimated to affect over 50 million Americans [38]. These disorders are associated with higher rates of driving and occupational accidents, and an increased risk of cardiac disease [70, 20]. Analysis of sleep is critical for the diagnosis, treatment, and scientific understanding of these disorders. Sleep analysis involves the identification of key properties in physiological recordings throughout the course of a night of sleep, particularly in the electroencephalogram (EEG). The EEG records the electrical activity of the brain, caused by the firing of millions of neurons, using electrodes placed on the scalp. EEG is the most commonly used signal in identifying the quality and progression of sleep, and often only a single electrode of EEG is necessary for sleep analysis.

Sleep is comprised of a number of stages, consisting of two main types, rapid eye movement (REM) and non-rapid eye movement (NREM). NREM can be divided further, into four stages (1, 2, 3, and 4) reflecting the continuum between drowsiness and deep sleep. This standard classification of sleep stages was devised by Rechtschaffen and Kales, and is referred to hereafter as the R&K system [58]. Sleep staging is the process of labeling each 30-second window in a recording with one of these stages. This is primarily done using the EEG, although other signals such as the electrooculogram (EOG) provide useful supplementary information. The set of annotations for an individual's entire night of sleep is referred to as a hypnogram. Normal sleep has a cyclic organization, in which individuals cycle from light to deep sleep, REM, and then return to light sleep. The transitions between states and the time spent in individual states carry information about the quality of sleep and insights into potential sleep disorders. The organization of sleep across the night is referred to as sleep architecture, and is evaluated using the hypnogram. The transitions between states and the time spent in individual states carry information about the quality of sleep and insights into potential sleep disorders.

Despite its essential role in sleep analysis, the current standard for sleep staging is commonly considered to have many problems. These derive from both the subjectivity of the sleep staging process and the simplifications inherent in the stage definitions, like discrete states and the failure to account for substantial inter-patient variability in the properties of sleep EEG. We address these limitations by learning topics over short patterns in the EEG, and using each window's mixture over topics as an alternative representation of sleep stages. There are several well known sporadically occurring patterns in sleep recordings, such as sleep spindles and k-complexes, indicating that a motif-based approach advantageous for modeling sleep EEG data. In contrast to most approaches to automated sleep analysis, this approach avoids the constrictive reliance on existing definitions of sleep stages by learning structure directly from the data and modeling mixtures of states. The unsupervised nature of the approach allows for the identification of new, richer structure in sleep data.

We learn these models on individual patients, as fitting a single model to a population may lose or misconstrue individual differences in sleep EEG properties. While universally defined sleep stages have a critical role in sleep analysis, the development of more expressive patient-specific models can provide complementary information that offers more detailed insights into a patient's sleep structure.

2.3.2 Limitations of Traditional Sleep Analysis

There are a number of problems with the traditional approach to sleep staging. First, R&K is often regarded as an oversimplification of the actual structure of sleep. The hard distinctions between stages, such as between stage 3 and stage 4 which reflect different levels of deep sleep, impose unnecessary structure on the data to facilitate manual annotation when the true progression towards deeper sleep is likely a continuous process. Additionally, the stages may not represent the full variety of sleep activity well. For instance, drowsiness (stage 1 sleep) has been classified into as many as 9 different stages [61], a level of detail which the R&K system is unable to capture.

Another issue with R&K is that inter-rater reliability (how well the annotations of two different experts match) is low, meaning that hypnograms of the same night of sleep from two different annotators may differ significantly. This detracts from the value of the R&K stages as a standard, as annotations from different sleep labs are not directly comparable. The subjectivity of applying the R&K standard has motivated the design of a wide variety of automated sleep staging algorithms [55]. Unfortunately, due to the subjectivity of the R&K standard, evaluating the accuracy of these stagers is difficult, as errors in labeling may actually reflect a reasonable alternative choice of annotation.

An unsupervised algorithm to identify sleep states could address both of these critical issues. An automatic method would give the consistency expected of a sleep staging standard, by avoiding the subjectivity of human annotators. By learning structure directly from the data, such an approach would also avoid the constrictive reliance on prior definitions of sleep stages, opening the door for richer descriptions and new ways to quantify sleep organization.

A third difficulty with the use of R&K is that the system was designed for young, healthy subjects. There are many individual differences in EEG and sleep organization [12], and applying a single set of staging definitions to a broad range of patients may pose difficulties in interpretation.

Prior work by Flexer et al. used a hidden Markov model (HMM) in an unsupervised approach to describe the structure of sleep recordings [27]. The method identified three population-wide states, corresponding to wake, sleep, and REM. The model used each window's posterior distribution over states as a continuous measure of sleep stages. A shortcoming of this approach is that although it results in a continuous mixture over states, the HMM assumes discrete states when estimating the model. This approach effectively simplifies R&K scoring even further, by reducing the set of states to three, limiting the model's ability to represent complex structure in the EEG. The approach also focused on modeling structure at the population level, rather than focusing on patient-specific modeling.

2.3.3 Methods

We explore the application of topic models as a method for unsupervised sleep analysis to develop a more nuanced description of the sleep EEG. In the context of sleep analysis, our goal is to identify latent states in the EEG recordings through "sleep topics" that can expand upon the information present in the sleep staging standard. The benefit of using "sleep topics" lies in LDA's assumption that a given data instance (a document) can derive from multiple topics, as opposed to a single state. This allows for model flexibility in identifying sets of potentially concurrent time-varying states, as well as allowing for states that relate to only a subset of features. We train models on individual patients, allowing each individual's sleep to be modeled separately. The development of patient-specific models avoids adverse effects inherent in fitting a single model to a population, where individual differences may be lost or misconstrued. While universally defined sleep stages have a critical role in sleep analysis, the development of more expressive patient-specific models can provide complementary information that expands the set of useful sleep analysis methods.

For each patient, the single-channel EEG recording was divided into non-overlapping one-second segments. From each segment we extracted spectral power in the four commonly used frequency ranges for EEG analysis (delta: <4 Hz, theta: 4-7 Hz, alpha: 8-13 Hz, and beta: 14-30 Hz). We discretized each of these features into five equiprobable bins on a per-patient basis using the SAX approach to time-series symbolization [48]. We treat these symbols as analogous to words, and apply Latent Dirichlet Allocation (LDA) to the resulting symbols. We divided each recording into non-overlapping 30-second segments, analogous to "documents". The use of 30-second segments was chosen so that the model would estimate topic mixtures for the same windows used in standard stage annotation. LDA was used to learn topics from the collection of 30-second windows, with a separate topic model learned for each patient. The number of topics for each patient was chosen using the Akaike Information Criterion. Parameter estimation was done with variational inference. The number of topics for each patient was chosen in a data-driven manner using the Akaike Information Criterion. The estimates for the posterior distribution of each window over the topics was used as a representation of the sleep recording, and was treated as an alternative to the R&K stages.

2.3.4 Evaluation

We evaluated the learned models to see whether the unsupervised model captured the same structure as the standard approach (that the learned models capture well-established properties of sleep structure) by using the topic mixtures for each window as features in a support vector machine (SVM) classifier for predicting the R&K stages. High accuracy for the SVM using topic mixtures as features would confirm that the model retains the relevant information from the R&K stages. This also illustrates a method to establish correspondence between the unsupervised model's structure and the currently accepted gold standard of sleep staging. To train and evaluate the SVM, the data was split into equally-sized training and test sets. Due to the patient-specificity of the topic models, predictors were trained on a per-patient basis. As a second point of evaluation, we assessed whether the topic models provided more information about the structure of sleep than the standard stages. Due to the difficulty of experimentally validating this property, we evaluate this by visualizing the resulting topic mixtures and qualitatively comparing them with the standard stages.

All evaluations were conducted on the publicly available MIT-BIH polysomnographic database from Physionet, consisting of a single channel of EEG from 15 patients with sleep apnea. Each recording contained a single channel of EEG, with recording durations ranging from 3 hours and 40 minutes to 6 and a half hours.

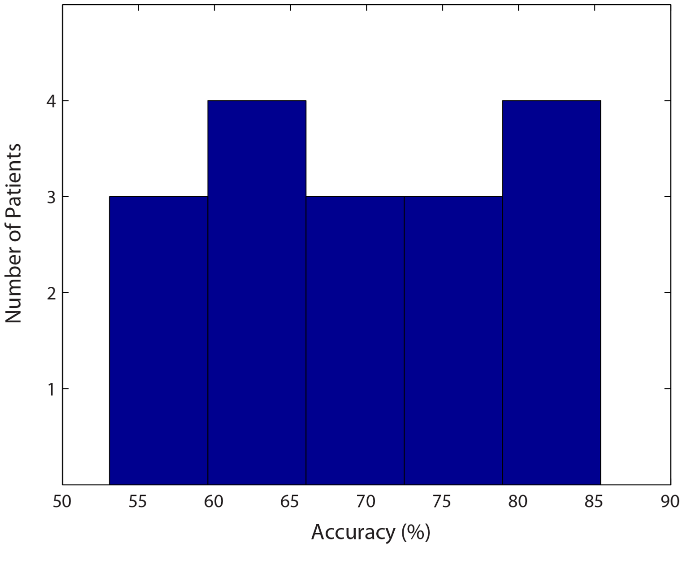

Figure 2.7: Accuracy of SVM models trained with topic mixture features when predicting R&K stages on the MIT-BIH dataset

Figure 2.7 shows a histogram of accuracies on the MIT-BIH dataset. The models achieved a mean accuracy of 70.1% and a median accuracy of 71.2%, indicating that the topic mixtures can predict R&K stages within the range of inter-rater reliability, which has been estimated between 70 to 90% [43]. Scores in this range by an automatic stager can be considered good performance as they achieve the same correspondence with the reference as another human annotator might. Despite the proposed method's discretization and unsupervised learning of structure, these results demonstrate that it preserves the information captured by the current standard of sleep staging.

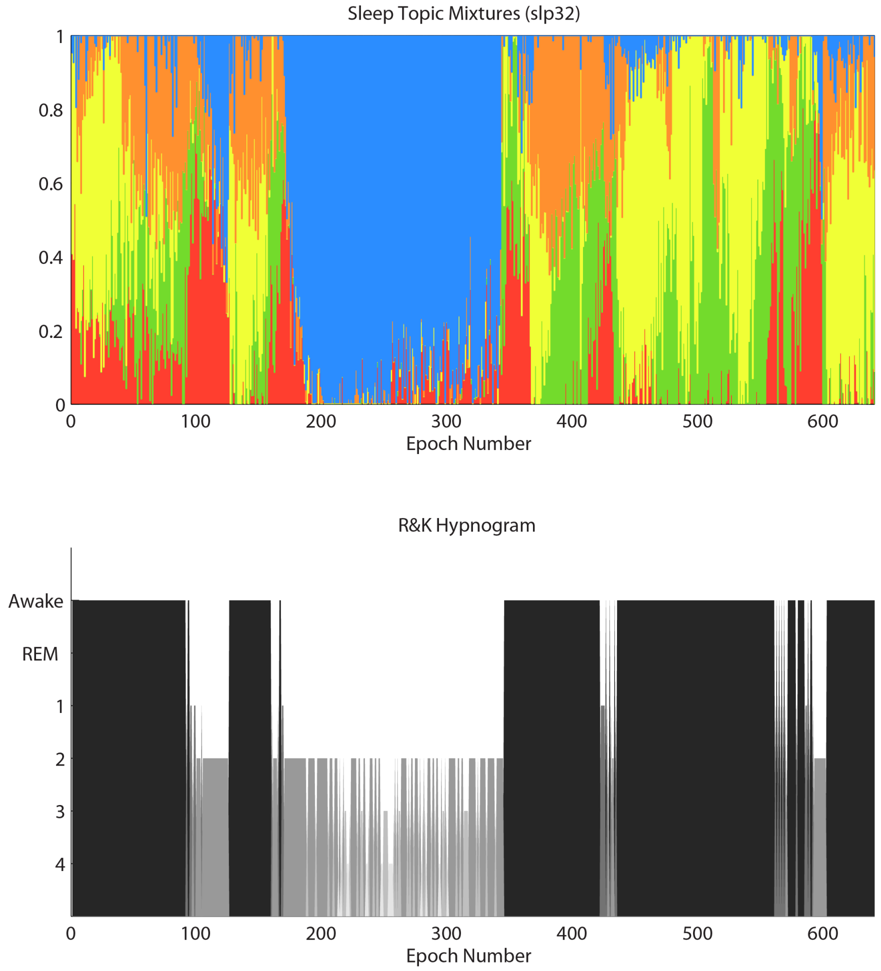

Figure 2.8: Topic mixture diagrams (top) and R&K hypnograms (bottom) for patient slp32.

Figures 2.8 and 2.9 indicate that the inferred models contain similar structure to the standard stages. The top panels depict the model estimates for each window's

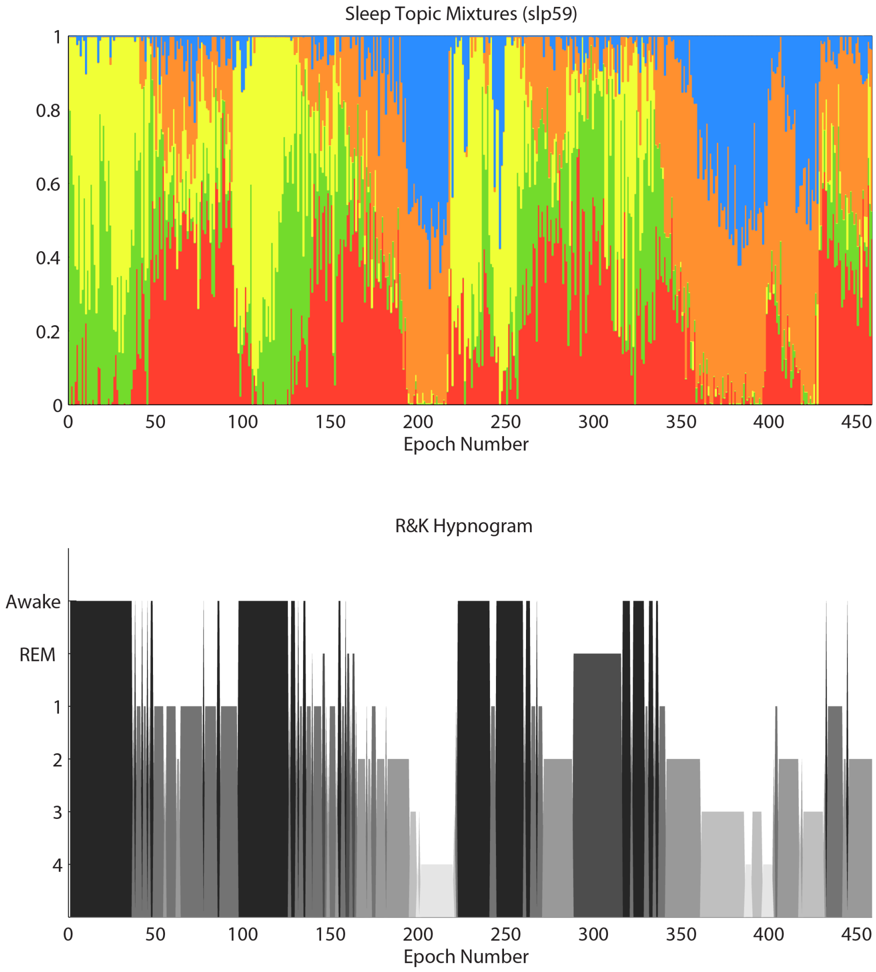

Figure 2.9: Topic mixture diagrams (top) and R&K hypnograms (bottom) for patient slp59.

distribution over topics, with each vertical strip corresponding to a 30 second time span. Each color in the mixture diagram represents a different topic. The size of a color band for a window indicates that topic's contribution to that window. The bottom panels show the standard annotations. For the patient in Fig. 2.8, the red topic (bottom) corresponds to light sleep, with increases as the patient enters stage 2 sleep and decreases as they begin entering stages 3 and 4. The blue topic (top) reflects deeper sleep (stages 3 and 4), dominating the recording from epochs 180 to 330. The patient in Fig. 2.9 reflects similar structure, with a topic corresponding to the awake state (yellow), another reflecting light sleep (red), and a third reflecting deeper sleep (blue).

Visual inspection reveals several properties that indicate greater expressive power than the sleep stages. The sleep topic model shows a continuous shift between states, reflected by gradual rather than abrupt transitions between different depths of sleep. This allows for finer assessment of the relative depth of sleep, for example, where a light sleep topic increases during waking epochs. The relative smoothness of transitions is a property of the data as opposed to the model, as the exchangeability of documents in the LDA model means that there is no explicit relationship between adjacent windows as in an HMM.

An observation with regard to the patient in Fig. 2.9 shows utility for sleep topics in going beyond the capabilities of R&K scoring. In the second patient's diagram, there are three long periods of stage 2 sleep, in epochs 160 to 200, 270 to 290, and 340 to 360. In the first and third instances, the patient progresses into deeper sleep (stages 3 and 4), while in the second instance they progress into REM. The topics capture this distinction well before the state transition, where the deeper sleep topic (top) is present and increasing in the first and third cases well before the transition begins, yet absent in the second. This indicates a difference in stage 2 sleep preceding the transitions, verifying the ability of the model to detect variations within a single

R&K stage.

2.3.5 Discussion

We presented an unsupervised approach to identifying structure in sleep EEG recordings. The sleep topic model improves on standard sleep analysis by providing an automated, data-driven algorithm for learning patient-specific sleep states. The approach relaxes the traditional assumption of discrete states, allowing for more expressive models that capture more detailed state information. The patient-specific approach allows for better modeling of individual differences, complementing the use of universal sleep stages. The resulting models are capable of representing standard sleep stages, while including additional information.

A clinical validation of the method by investigating the derived topics or the relationships between the derived models and sleep disorders is an area for future work. A possible extension to the presented methods could incorporate additional features, from other time-scales or additional physiological signals.

CHAPTER III

Learning Risk Factors from Long-Term ECG

3.1 Introduction

Physiological recordings have unutilized predictive information that can improve the accuracy of clinical risk models, and the development of new risk factors is the key to leveraging this data for better patient risk estimates. However manual identification of high-risk properties in the data is challenging due to the subtlety of changes, the large quantity of data, the large space of possible hypotheses, and limited clinical understanding of the relationship between physiology over time and patient risk. Better computational approaches to feature learning are the most promising tool for capturing the predictive information in long-term physiological recordings for inclusion in risk models.

The limited prior work on learning risk factors from these recordings is based on measuring or combining predefined features on the data. The work discussed in Chapter 2 on heart rate topic models also relies on patterns defined using a discretization scheme, where the learning process identifies structure in the occurrence of these patterns rather than learning the patterns themselves. In this chapter we investigate a powerful class of methods, convolutional neural networks (CNNs), to learn features from physiological recordings for clinical risk stratification. We investigate CNNs specifically for learning cardiac risk factors from long-term heart rate recordings, and demonstrate that these computationally-derived features result in more accurate estimates of risk than prior approaches. Several studies have applied CNNs to classification tasks on physiological time series [53], however they have focused on labeling tasks and not the prediction of patient outcomes. To our knowledge, this is the first investigation of CNNs for learning features from physiological recordings for the prediction of patient outcomes.

3.1.1 Long-Term Heart Rate for Prediction of Cardiovascular Death

Patients are at an elevated risk of death from cardiovascular events following an acute coronary syndrome (ACS). Risk of future events varies widely among patients following an ACS [21], and there are a wide range of treatments available. Accurate risk stratification is essential for matching patients to the appropriate level of treatment.

Long-term heart rate information has been shown to improve patient risk assessment for cardiovascular death (CVD) when included alongside standard clinical and echocardiographic risk factors [5, 68]. The use of long physiological recordings provides longitudinal information about the functioning of the heart that traditional risk factors fail to capture. Costs for recording and storing large quantities of heart rate data continue to drop, permitting the collection of more data from more patients. This provides large quantities of data to leverage for the development of more effective risk factors, and improves the feasibility of incorporating long-term heart rate into the decision making of standard clinical practice.

A variety of heart rate risk factors have been identified, each designed to measure the occurrence of a particular kind of pattern associated with a higher risk of adverse events. Heart rate turbulence models heart rate changes following premature ventricular contractions [4], deceleration capacity models changes surrounding decelerations in the heart rate [6], and heart rate variability metrics consider overall variation across a recording [11]. The diversity and complementarity of these measures indicates the wealth of predictive information in long-term heart rate, however each measure targets only one of the many predictive aspects of the heart rate. This limits the utility of any one measure, and it is unclear which subset of these factors, or what method of combining them, is most effective for risk stratification. Additionally, it is unclear how to modify these measures to optimize their ability to risk stratify patients from different populations or with respect to different clinical outcomes than those for which they were originally proposed.

In contrast to the expert guided, hypothesis-driven development of traditional heart rate risk factors, computational methods automatically develop risk markers from large collections of physiological recordings. These methods identify many different patterns in the heart rate associated with adverse outcomes in a systematic and data-driven way, and incorporate them into a single predictive measure. Where traditional measures each target a particular property of the heart rate, these methods can generate a single multi-faceted estimate of risk, and can more easily develop risk factors optimized for different outcomes and patient populations. Several prior works have proposed computational risk scores for long-term heart rate recordings [68, 23]. However, the effectiveness of these approaches is limited by their requirements for coarse data discretization and their restrictive definitions on the kinds of patterns that they can learn.

In this chapter, we investigate a more flexible and powerful computational approach to learning a risk stratification measure directly from collections of heart rate recordings. We make use of convolutional neural networks (CNNs), responsible for recent advances in the areas of computer vision and speech recognition, to automatically identify a set of characteristic patterns for risk stratification from large amounts of Holter-derived heart rate data. We develop a risk score using a CNN for the prediction of CVD post-ACS in a large and well-characterized population of patients presenting with non-ST-elevation ACS. We compare supervised and unsupervised approaches, and consider a variety of network input features derived from the ECG. When developed using heart rate recordings as inputs, we find that a CNN risk score not only predicted adverse outcomes more accurately than previously proposed heart rate metrics, but improved overall risk stratification when combined with standard clinical risk measures.

3.2 Background

3.2.1 Prior Work on Cardiac Risk Stratification from Long-Term ECG

The vast majority of approaches to feature extraction from ECG recordings first extract low-level features for each beat or pair of adjacent beats in the recording, and then compute some high-level features on this derived time series. ECG feature extraction methods can be categorized based on which low-level features of the ECG they use: many focus on the time between adjacent beats (heart rate), while others focus on differences in morphology between adjacent beats.

3.2.1.1 Expert-Designed Heart Rate Measures

Features designed to capture structure in heart rate are the most commonly studied. Many measures have been proposed by experts to identify high-risk properties in long-term heart rate recordings.

Heart rate variability (HRV) metrics have been investigated extensively for decades, and a variety of approaches to its measurement have been shown to be related to a range of adverse outcomes [11]. Useful time-domain measures of HRV include computing the standard deviation of NN-intervals in 5 minute windows and averaging across all windows, or alternatively taking the average NN-interval in all 5 minute windows and taking the standard deviation of these averages over the recording.

Looking at smaller windows is important, as simply taking the standard deviation of all NN-intervals over the day is not useful for prediction, especially considering the large cyclic variation in heart rate over the course of a day. Frequency-domain measures have also been proposed, one of the most useful of which is the ratio between power in low (0.04 to 0.15 Hz) and high (0.15 to 0.40 Hz) frequency intervals.

More recently, measures that target more specific patterns of variability have been associated with adverse otucomes. Heart rate turbulence (HRT) [4] looks at heart rate changes immediately before and after premature ventricular contractions (PVCs). There are two components proposed to measure heart rate turbulence. Heart rate turbulence onset compares the average NN-intervals between the two beats before and after a PVC, measuring the change in heart rate immediately following the PVC. Heart rate turbulence slope averages the first fifteen NN-intervals following all PVCs in the dataset, and identifies the maximum regression slope using a 5 NN-interval long sliding window. This identifies the fastest average acceleration in heart rate in the periods immediately following PVCs. Deceleration capacity (DC) [6] incorporates information only around decelerations of the heart rate. More specifically, DC identifies all decelerations (when an NN-interval is longer than the one that preceded it), averages the signals in windows around these decelerations, and measures the change in heart rate in the intervals immediately before and after the deceleration. While HRT and DC have each been found to be predictive of adverse outcomes, they have shown even better predictive power when combined into a single risk factor [5].

Measures designed by experts to capture high-risk structure may not be the most effective way to quantify the underlying physiological phenomena that they aim to measure. Additionally, because these measures are identified in a hypothesis-driven way, there may be additional kinds of high-risk structure in heart rate that have not been investigated. Systematic, data-driven approaches hold the promise of identifying predictive structure in a more effective and less biased way. They also have the potential to simplify the space of long-term ECG risk factors. The large number of complementary features makes it more challenging to include long-term ECG features in risk stratification models, as the appropriate subset and combination of these features may vary depending on the population and outcome of interest. A single multi-faceted and automatically generated risk score could alleviate these problems, and make the incorporation of long-term heart rate into clinical decision making more practical and effective.

3.2.1.2 Morphology-Based Measures

In addition to features derived from the heart rate, several methods have been designed to capture patterns in morphological properties (i.e. the shape) of the ECG. Microvolt T-wave alternans measures beat-to-beat fluctuations in the amplitude of the T-wave, and has been found useful for prediction of a wide range of cardiac events [54]. Morphologic variability (MV) is a broader approach, that measures beat-to-beat changes in morphology over the entire beat, not just the T-wave. MV also takes a more computational approach, using dynamic time warping to measure small morphological differences between adjacent beats, using the Lomb-Scargle periodogram to measure the spectral properties of the morphologic distances over five minute windows, and then measuring the 90th percentile of the energy in a particular frequency range. Several versions of MV have been found to be predictive of adverse outcomes [68, 50].

3.2.1.3 Computational Measures

Several works have investigated automatic methods to learn predictive structure in heart rate recordings. The method of heart rate motifs identifies heart rate patterns lasting several beats (less than 15) that are associated with higher or lower risk of events [68]. In contrast to measures like DC and HRT, which have been designed by experts to capture very specific patterns of variability in the heart rate, this approach has the flexibility to identify and aggregate many such patterns into a single model. As discussed in Chapter 2, learning higher-level structure in the cooccurrence of these motifs using topic models can identify additional information that improves prediction of adverse cardiac events. Yet while these approaches have demonstrated that it is possible to improve prediction of adverse events using computationally-derived scores, their methodological constraints have limited their predictive power. A challenge with the use of heart rate motifs and topics is that they necessitate a coarse discretization of the heart rate. While work has been done to counteract this oversimplification of the data by considering approximate pattern matching [18], this discretization still loses many subtle changes in the heart rate and makes it computationally infeasible to incorporate motifs of greater lengths.

3.2.2 Convolutional Neural Networks

Convolutional neural networks (CNNs) are neural networks that learn convolutional feature detectors from a collection of data to minimize some objective function. Unlike standard neural networks, where the lowest layer has parameters associated with all values of the input data (global feature detectors), CNNs learn features associated with small patches of the input data that are applied repeatedly across the entire input data. This improves generalizability by learning feature detectors that can identify patterns regardless of where in the input data they occur, and greatly reduces the number of parameters needed to model large inputs.

CNNs contain several kinds of layers. Convolutional layers slide a set of convolutional filters (i.e. feature detectors) across the entirety of their input, resulting in a feature activation for each filter at each position of the input. The filters used in the convolutional layers are not prespecified, and are learned from the training data. Pooling layers aggregate their input values over fixed size regions. Common functions used for pooling are the maximum or the mean. Pooling layers are useful for reducing the size of their input data, thereby reducing the size of subsequent layers in the network. This reduces the number of parameters needed at higher layers, and allows those higher layers to process the data at increasing scales: while a feature detector at a lower layer may span a very small region of the input, a feature detector at a higher layer may identify structure over larger patches or even the entire input data. The ability to learn structure at multiple scales is a major reason for the success of CNNs in a variety of application domains. Inner product layers, which are fully connected, are generally used as the highest layer(s) of a network to combine the convolutional feature activations to minimize the loss function. Common choices for loss are softmax, hinge, and euclidean distance (for regression). Non-linear transformations are often applied after convolutional or inner product layers. Rectified linear units (which map all negative values to zero), sigmoid functions, or the hyperbolic tangent are the usual choices. Parameters of the network are usually learned using backpropagation and some variant of stochastic gradient descent.

Standard CNNs are supervised models. There are several unsupervised alternatives: stacked autoencoders and deep belief networks. Autoencoders learn a feature mapping that can be used to "encode" the input data, and can be inverted to "decode" encoded data to approximate the original data. Autoencoder parameters are learned to minimize the reconstruction error when training data is encoded and decoded using the parameters. Optimizing autoencoders with multiple layers using traditional methods is challenging, as gradients from the error shrink quickly as the number of layers increases. To address this, training for these models is usually done in a greedy layer-wise fashion: the first layer is learned on its own, and then the second layer is learned after fixing these first layer parameters.

Deep belief networks (DBN) are another unsupervised neural network variant. Each layer of a DBN is a restricted Boltzmann machine (RBM). RBMs are 2-layer undirected graphical models, with the connections between layers represented by a weight matrix. The RBM parameterizes the joint distribution of the input data and the hidden units, with the probability density function defined as the exponential of an energy function. With real-valued variables, the visible units are distributed under a Gaussian with diagonal covariance. The objective when learning the parameters of an RBM is to maximize the training data likelihood. Estimating the parameters using stochastic gradient descent is intractable because of the need to compute the exact gradient of the data log likelihood, so an alternative method called contrastive divergence is used [37]. DBNs, like stacked autoencoders, are usually trained in a greedy layer-wise manner. Probabilistic maximum pooling has been developed for use with convolutional DBNs [45], enabling them to incorporate pooling over input regions. A main advantage of using a DBN instead of a stacked autoencoder or a traditional CNN is that a DBN is a generative model, and can be used to generate new data points.

Both supervised and unsupervised approaches to developing CNNs have been found effective for a variety of problems, including classification of time series data [46]. CNNs have proven very successful in many tasks, where the learned features have substantially outperformed previous state-of-the-art methods [7].

3.2.3 CNNs for Cardiac Risk Stratification

We propose the use of CNNs to learn predictive features from long-term ECG recordings. Unlike prior methods for learning from this data, CNNs use a more powerful and flexible representation that can identify structure in these recordings: they avoid discretization of inputs, can learn patterns at a variety of time scales, and can allow integration of multiple ECG parameters into a single model (e.g. heart rate and pairwise morphologic distances). Learning hierarchical structure is a major advantage of deep learning, as it can capture patterns on the order of several beats to several minutes to several hours. Hierarchical structure was shown in Chapter

2 to add value when learning a topic representation over heart rate motifs, and it likely that ability to extract higher-level features on top of lower-level ones will aid in prediction. Unsupervised methods for learning CNNs derive task-independent features, which is particularly useful given the range of clinically interesting cardiac outcomes. Alternatively, there are also various kinds of possible supervision: long-term prediction (SCD or CVD in the next few years) and short-term prediction and detection (transient ischemia, ventricular tachycardia and pauses) labels are available for many patients.

There were a number of questions we aimed to address with this investigation:

- What are the most useful low-level features (e.g. raw heart rate, raw pairwise morphologic distances, spectral features) for CVD prediction, and is combining them useful for prediction?
- How should these features be normalized, in a patient-specific or population-wide manner?
- Is a supervised or unsupervised approach more effective for this task?

3.3 Methods

3.3.1 Study Population

Data from the Metabolic Efficiency with Ranolazine for Less Ischemia in Non-ST-Elevation Acute Coronary Syndrome (MERLIN) -Thrombolysis in Myocardial Infarction (TIMI) 36 trial were used for develop the CNN models. The MERLIN-TIMI36 trial compared the efficacy of Ranolazine, an anti-anginal drug, to matching placebo. The study enrolled 6,560 patients within 48 hours of admission for non-ST-elevation ACS, and patients received standard medical and interventional therapy according to local practice guidelines. Patients were followed for up to 2 years, with a median follow-up of 348 days. Full details on the study have been reported elsewhere. Patients had 24 hours of continuous Holter recordings, from which heart rate was extracted. Patients were excluded from the analysis if Holter recordings were unavailable, resulting in a set of 5,117 patients available for model development. Patients were excluded from the final analysis if their recordings were of insufficient signal quality to compute all of the heart rate risk factors, or if no LVEF measurement was available. This resulted in a final set of 2,599 patients available for evaluation.

Data from the DISPERSE2 trial were used as a validation set to select model hyperparameters. The DISPERSE2 trial compared the safety and efficacy of AZD6140 with clopidogrel, and enrolled 990 patients within 48 hours of admission for non-ST-elevation ACS. Patients received standard medical and interventional treatment according to local practice guidelines, and were followed for up to 90 days. Full details on the study have been reported elsewhere [15]. Patients had 24 hours of continuous Holter recordings, from which heart rate was extracted. Patients were excluded from the analysis if their recordings were of insufficient signal quality, resulting in a final set of 863 patients.

3.3.2 Comparison of Different ECG Feature Sets and Cardiac Outcomes

Using features learned by a DBN, we compared a variety of feature sets for the prediction of several cardiac outcomes. The feature sets considered included: raw NN-intervals, raw morphologic distances, 5 minute periodograms of NN-intervals, 5 minute periodograms of morphologic distances, and the average and standard deviation of NN-intervals in 5 minute windows. Several other combinations were considered when training the model: raw NN-intervals and morphologic distances, and 5 minute periodograms of NN-intervals and morphologic distances. We also considered combining features for these combinations when models were learned for each feature set separately. Hyperparameters were chosen separately for each feature set using a grid

	CVD	SCD	MI
NN	.696	.572	.556
MD	.531	.550	.392
NN and MD	.681	.607	.542
NN Periodograms	.675	.515	.554
MD Periodograms	.620	.523	.526
NN and MD Periodograms	.637	.501	.500
NN Window Stats	.599	.577	.534

Table 3.1: AUC values for three outcomes on the MERLIN dataset using RBMs trained with various ECG feature sets. NN - NN-intervals, MD - morphologic distance, Window Stats refers to mean and standard deviation for each window.

search. Results shown were generating using only one layer, as experiments with two layers indicated that the only feature set that benefited from an additional layer were the raw NN-intervals.

Models were developed for each feature set normalizing the input data two ways, either patient-specific or population-wide. Normalization consisted of shifting and scaling the data to have zero mean and unit variance. The raw interval features did better with patient-specific normalization, suggesting that it was relative changes in their values that indicated risk of events. The remaining features, computed over 5 minute windows, had better prediction with population-wide normalization and poor prediction with patient-specific normalization, suggesting that it is the difference in feature values between patients (as opposed to differences in feature values over time relative to a patient's baseline) that is useful for prediction.

Raw NN-intervals achieved the highest AUC for CVD prediction, followed by NN periodograms. Combining the two feature sets did not improve prediction over just using raw NN-intervals. The raw morphologic distance features were not predictive of any of the outcomes. This is likely due to the high levels of noise in the time series, which is why the morphologic distance periodogram features did better for CVD prediction. The average and standard deviations of 5 minute windows were most useful for SCD prediction, along with the raw NN-intervals, although the AUCs were much lower across all feature sets than they were for CVD. None of the feature sets were useful for prediction MI. Due to the lack of complementarity seen between feature sets, the presence of an improvement using multiple layers, and the superiority in CVD prediction, raw NN-intervals were the focus of subsequent experiments.

3.3.3 Unsupervised Approach

Convolutional DBNs were used as the unsupervised network in the evaluation. The first 24 hours of heart rate data for each patient were used as inputs to the DBN to learn the convolutional filters. Evaluation was done by computing the hidden unit activations for a heart rate recording, and aggregating the pooled activations from each hidden layer. This aggregation consisted of the mean and standard deviation of each filter's activations across the recording. We considered adding additional features such as the median, max, and 90th percentile, but they did not improve prediction.

An l_1-regularized logistic regression model was trained using these aggregated feature activations to predict CVD. Cost-sensitive weighting was used to increase the cost of misclassifying the less frequent positive examples, which resulted in substantial improvements in discrimination. An SVM with a radial basis function kernel did not improve prediction, and neither did the use of an AUC-maximizing formulation of the SVM. Evaluation was done using leave one out cross-validation. Data from the DISPERSE2 trial was used to select the model hyperparameters (classifier regularization, number of convolutional filters, filter length, and pooling factor).

The final network architecture used in the evaluation is shown in Fig. 3.1.

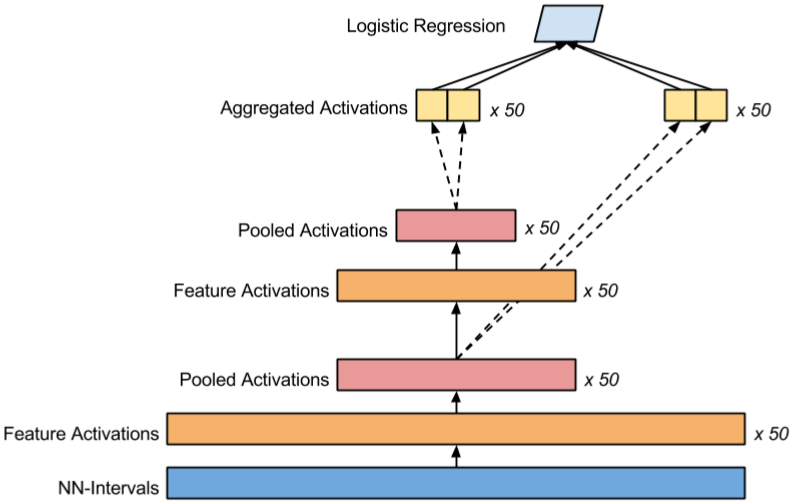

Figure 3.1: Unsupervised architecture used in the experiments. 24 hours of NN-intervals were used as input to the network. Two sets of convolutional layers with 50 filters of length 20 were each combined with maximum pooling layers, which reduced the size of each layer by a factor of 5. After the network was trained, the average and standard deviation of each maximum pooling layers activations were used as predictive variables in a logistic regression model trained to predict CVD.

3.3.4 Supervised Approach

Caffe was used to train and test the supervised CNNs [39]. The first 20,000 beats of a patient's recording were used as inputs to the network. Experiments found that this was a better choice than 10,000 or 50,000 beats. The inclusion of multiple input patches from each patient's recording was considered to increase the number of training data points, expand coverage of the data, and reduce the chances of missing high-risk phenomena by selection of a subset of data. These experiments, which considered both sliding and non-overlapping windows, used all windows from a patient as training examples, and during evaluation predictions for all of patients windows were aggregated (using the mean or max prediction across all windows) to generate a single prediction. The use of multiple windows per patient did not affect the discrimination of the model, and substantially increased the amount of time needed to train the models, and so the remaining experiments used only a single 20,000 beat sequence from a patient's recording when developing and applying the CNN.

Due to the rarity of positive examples (CVD occurred with a rate of 4% in MERLIN), positive examples were oversampled to improve the balance between positive and negative examples. A cost-sensitive hinge loss function was tested to see if improvements in discrimination similar to the unsupervised case were possible, however the cost-sensitive loss did not noticeably affect the model's predictions.

The networks were trained and evaluated with a five-fold cross-validation scheme. Hyperparameters selected included the number of convolutional filters, filter length, and pooling factors. A softmax loss function was used in the networks, as it resulted in better discrimination than when the network was trained using hinge loss.

A number of variations to the network architecture were considered, but were not used due to a lack of improvement over simpler alternatives. Multiple inner product layers led to overfitting, even when using rectified linear units and/or dropout. Rectified linear units after the convolutional layers hurt prediction. Other loss functions (aside from loss when predicting CVD) were considered as additional signals to help learn more useful features. These included other outcomes (myocardial infarction, sudden cardiac death) as well as previously proposed heart rate risk factors (heart rate turbulence, deceleration capacity). After experimenting with a range of weighting schemes relative to the primary loss (CVD), the use of additional losses did not improve CVD prediction, although the inclusion of existing heart rate risk factors as secondary losses increased the correlation of the CNN risk score with the risk factors included.

Including previously proposed heart rate risk factors (deceleration capacity, heart rate turbulence, heart rate variability) as additional inputs to the inner product layer was considered, with the hopes that this would encourage the network to identify features that were complementary to the known risk factors, and might improve overall prediction when the CNN score was combined with the existing risk factors. Experiments found that while this approach did make the CNN score more complementary to existing risk factors (correlations with these risk factors dropped), it hurt the utility of the CNN score by itself (AUC=0.70 vs. 0.71), and it did not affect prediction using the CNN score with the existing risk factors (AUC=0.72 vs. 0.72).

We also considered pooling not just over time, but over groups of filters within a layer. This could reduce redundancy in features learned by grouping activations of similar or shifted versions of a pattern. Experiments over a variety of pooling schemes found no improvement over the simpler baseline.

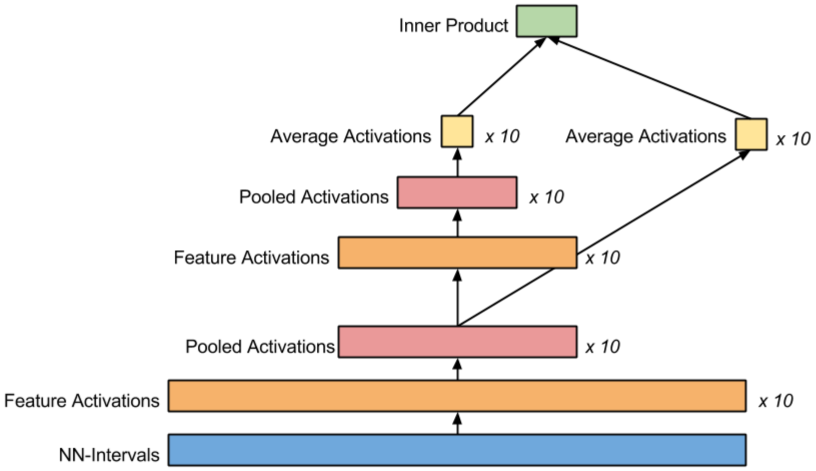

Figure 3.2: Supervised architecture used in the experiments. A sequence of 20,000 NN-intervals was used as input to the network. Two sets of convolutional layers with 10 filters of length 20 were each combined with maximum pooling layers, which reduced the size of each layer by a factor of 20. Two global average pooling layers calculated the average activations of each filter from the maximum pooling layers. These average activations were inputs to an inner product layer, optimized to minimize the softmax loss for CVD prediction.

The final network architecture used in the evaluation is shown in Fig. 3.2.

3.4 Results

3.4.1 Filters Learned

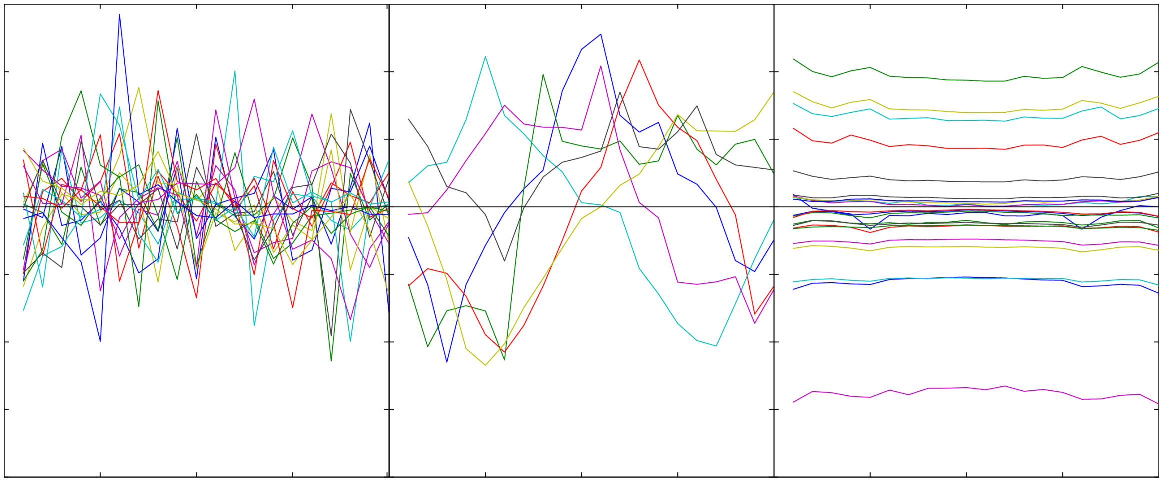

Figure 3.3: Three clusters of first layer filters learned by the convolutional DBN.

Fig. 3.3 shows three clusters of filters learned by the convolutional DBN. These clusters correspond to high variability (left), gradual accelerations/decelerations (middle), and flat filters (right). The high variability filter activations correlated positively with heart rate variability and heart rate motifs. The gradual accelerations/decelerations correlated negatively with deceleration capacity, heart rate turbulence, and heart rate variability. The flat filters were negatively correlated with heart rate variability and heart rate motifs.

These correlations between the filter activations and existing heart rate measures resulted in higher correlations between the existing measures and the final DBN score. Out of all of the heart rate measures evaluated, the DBN risk score was the most highly correlated with deceleration capacity, heart rate turbulence onset, heart rate variability, and heart rate motifs, and was the second most correlated measure with respect to heart rate turbulence slope. This indicates that the diversity of features learned by the model allowed it to capture a larger range of high-risk behavior than existing measures, even compared to heart rate motifs.

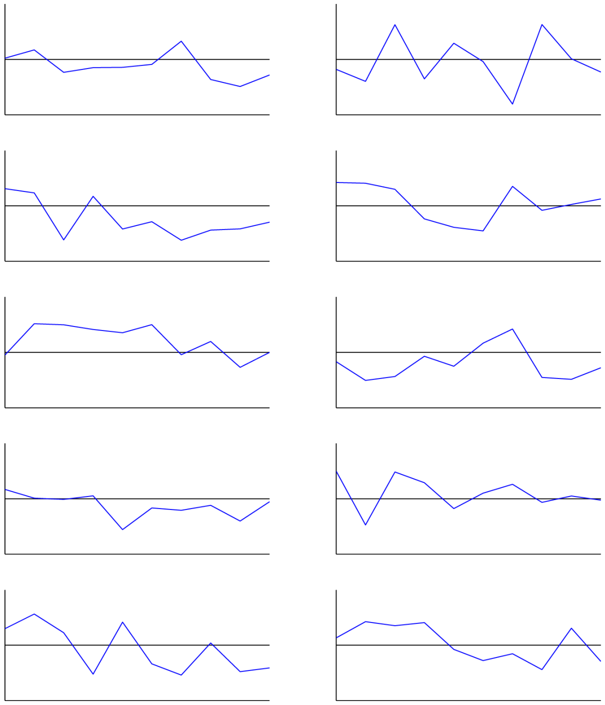

Figure 3.4: Ten first layer filters learned by the CNN.

Fig. 3.4 shows the first layer filters learned by the CNN. The optimal number of filters was much smaller than for the DBN.

3.4.2 Comparison to Heart Rate and Clinical Factors

The area under the ROC curve for the heart rate risk factors, and p values indicating the significance of an improvement of the proposed score over existing risk factors are shown in Table 3.2. The supervised and unsupervised CNN scores had higher AUCs and hazard ratios than the other heart rate risk factors, including the topics score (AUC: 0.725 and 0.694 respectively), although the improvement was not always statistically significant.The supervised score had a higher AUC and hazard ratio than the unsupervised score, although this difference was not significant. Both CNN scores had significantly higher AUCs than the TRS (p=0.044 and 0.005), as well as higher hazard ratios. The supervised CNN score had a larger AUC than LVEF, although

	AUC	p (vs. Unsup.)	p (vs. Super.)	Hazard Ratio	Adjusted Hazard Ratio
TRS	.672	.044	.005	2.7	n/a
LVEF	.708	.623	.294	5.3	n/a
DC	.667	.228	.050	2.9	2.1
HRT1	.619	.018	.001	1.8	1.6
HRT2	.646	.087	.011	2.2	1.4*
HRV	.669	.242	.055	2.8	2.1
Motifs	.659	.161	.029	2.4	1.9
Topics	.685	.402	.122	2.9	2.2
Unsupervised	.694	Ref.	.181	3.5	2.4
Supervised	.725	.819	Ref.	3.6	2.6

Table 3.2: Area under the ROC curve (AUC) and p values indicating the significance of an improvement of the proposed score over existing risk factors, as well as univariate hazard ratios and hazard ratios after adjusting for TRS and LVEF.

the improvement was not statistically significant. All risk factors had statistically significant hazard ratios when evaluated individually, and all adjusted hazard ratios were also significant, except for heart rate turbulence slope.

Fig. 3.6 shows the ROC curves for the five risk scores with the highest AUCs: LVEF, TRS, the topics score, and both CNN scores. LVEF dominates the other measures in the low sensitivity/high specificity part of the curve. The supervised score dominates the other measures in the medium to high sensitivity/medium specificity section of the curve. The supervised and unsupervised score ROC curves nearly dominate the TRS curve, and the supervised score also dominates the topic ROC curve. The curves suggest that while LVEF is still the clear best choice for high specificity thresholds, the supervised CNN score is more effective than the other measures at medium to low specificity thresholds.

Fig. 3.6 shows the AUCs at each step of a logistic regression backwards stepwise elimination process using all clinical and heart rate measures as inputs. There was a large amount of redundancy seen in the heart rate features, with no virtually no drop

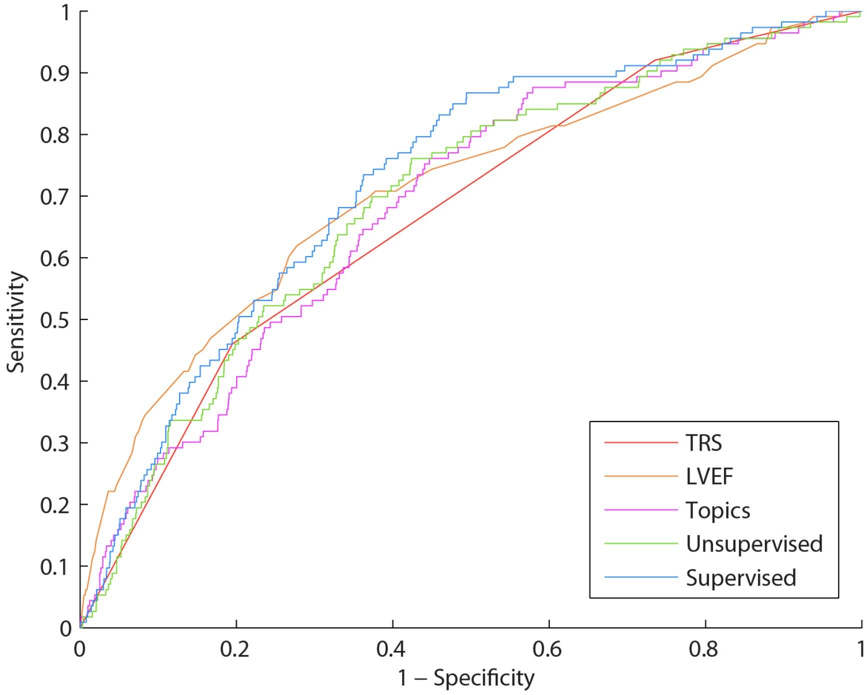

Figure 3.5: ROC curves for LVEF, TRS, Heart Rate Topics, and both CNN scores.

in the AUC after removing 5 of the features. The unsupervised score was removed early, due to its redundancy with the supervised score and its lower discrimination. The only heart rate measures selected were the topics score and the supervised score. TRS was removed before the supervised score. LVEF was the most significant variable in the model.

Table 3.3 shows statistics comparing the logistic regression model containing all previously proposed risk factors with a model that additionally contains the CNN risk scores. The addition of the supervised score resulted in a larger improvement in AUC relative to the addition of the unsupervised score (0.793 vs. 0.784), although neither improvement was statistically significant. The supervised score achieved statistically significant improvements in IDI, NRI, and category-free NRI.

Inclusion of the supervised score also resulted in a larger improvement in the hazard ratio than inclusion of the unsupervised score (HR: 6.5 vs. 5.5). The baseline

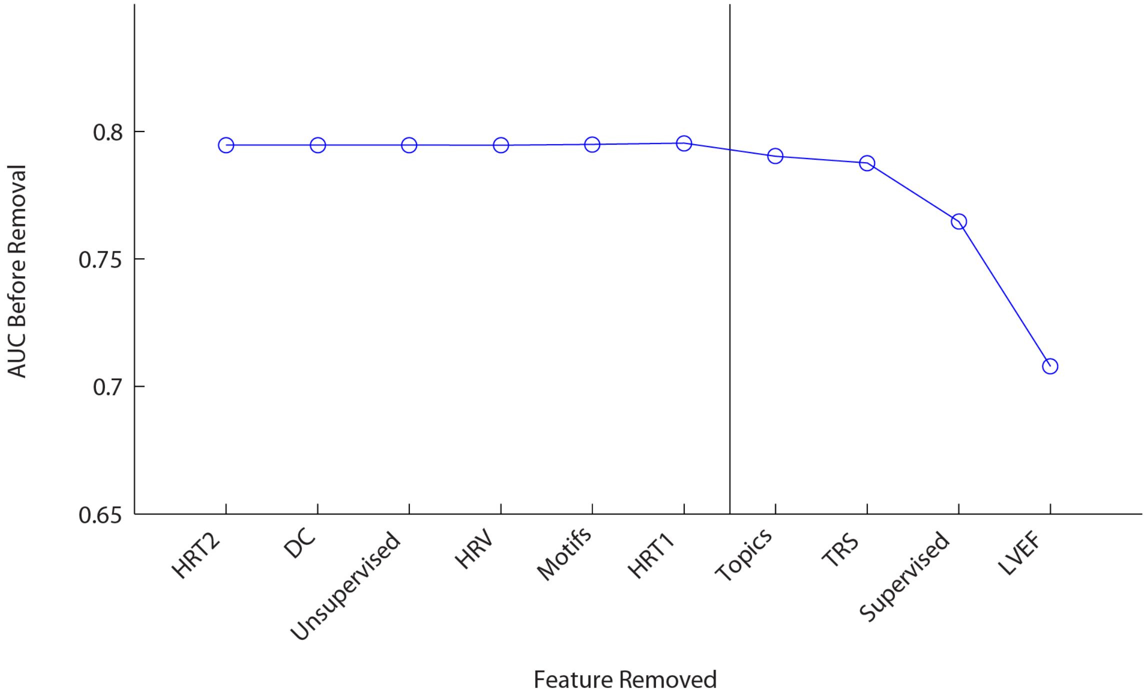

Figure 3.6: Variables removed and AUC of the model before removal for a backwards stepwise elimination model including all risk scores. The vertical black line indicates what variables were selected using a p<0.05 cutoff.

	Baseline	with Unsup.	with Super.
AUC	.779 (Ref.)	.784 (.441)	.793 (.330)
IDI	Ref.	.002 (.108)	.005 (.049)
NRI	Ref.	.019 (.201)	.056 (.025)
Category-free NRI	Ref.	.076 (.213)	.182 (.029)
HLχ^2	6.51 (.590)	4.62 (.797)	10.03 (.263)
Hazard Ratio	5.08 (<.001)	5.49 (<.001)	6.5 (<.001)

Table 3.3: Comparison of a model using the existing heart rate derived risk factors with models that additionally include either the unsupervised or supervised CNN scores. Area under the ROC curve (AUC) and p value reflecting improvement relative to the baseline, integrated discrimination improvement (IDI), net reclassification improvement (NRI), category-free NRI showing improvement relative to the baseline, Hosmer-Lemeshow χ^2 score, and hazard ratio are shown.

model and both models including CNN scores were well calibrated according to the Hosmer-Lemeshow chi-squared score ($p > 0.05$).

3.5 Discussion

This chapter discussed the use of CNNs for automatically developing a risk factor from a collection of heart rate recordings. When evaluated on a large cohort of patients admitted for non-ST-elevation ACS, the CNN score more accurately predicted CVD than previously proposed heart rate risk factors, including both expert-designed and computationally derived measures. The proposed score not only achieved superior discrimination of CVD, but identified novel, independent information that improved overall risk stratification when combined with the set of existing risk factors.

The supervised CNN consistently outperformed the unsupervised CNN, indicating that the use of patient outcomes when learning the model improved its ability to distinguish patients likely to suffer CVD following an ACS.

We proposed a general purpose approach to learning a heart rate risk factor. While the model learned and evaluated in this work is specific to risk stratifying patients for CVD following a non-ST-elevation ACS, the approach can be used with different datasets to develop risk factors tailored to different populations and clinical outcomes.

There were several limitations to the analysis. There was a median follow-up of one year in the MERLIN-TIMI36 cohort (maximum of two years), and additional patients may have suffered CVD after the duration of the study. The study was retrospective, and a prospective evaluation on a large patient population is necessary to confirm the utility of the CNN risk scores.

There are a number of directions for future investigation. In the experiments conducted there was no value found in combining heart rate and morphologic distance features into a single CNN architecture, or even in combining the features learned by separate CNNs in a final model. While this may suggest that morphologic distance features capture much of the same information as heart rate, it is possible that alternative approaches to incorporating morphologic distance into models will results in improvements. Alternatively, there may be other morphologic features (e.g.

QT intervals, magnitudes of individual waves) that add value when included either individually or alongside heart rate as inputs to a CNN.

CHAPTER IV

Optimizing Imputation for Classification with Missing Values

4.1 Introduction

Missing values are abundant in real-world medical data for a variety of reasons: sensors may malfunction, equipment may be unavailable, or a patient may not be in a condition where certain measurements can be recorded. This is a barrier to risk stratification, as commonly used multivariate risk scores are not designed for use with missing covariates. Risk models that are robust to the presence of missing data would have the ability to risk stratify the large percentage of patients in real-world settings that do not have observed data for all model variables. Better handling of incomplete data would also allow the inclusion of additional risk factors that are currently excluded from the development of risk models due to their limited availability in standard practice, improving the quality of predictions when these additional variables are available. Unfortunately standard classification models, including logistic regression (the most commonly used method to develop risk models in clinical domains), are unable to handle missing data in either the training or testing phases. While some classification models can readily handle missing values (e.g. random forests, k-nearest neighbors), these approaches do not result in models that can be easily understood or validated by the clinical community.

Developing and evaluating risk models on patients with missing values is generally handled by imputation of the missing observations. Recently, studies in the machine learning literature have investigated the handling of missing values specifically in the context of classification. These methods can be organized into several different categories. Some methods eschew imputation models entirely, attempting to learn classifiers that can robustly handle inputs with missing values without attempting to fill them in [17, 32]. Others learn an imputation model using standard methods (such as the EM algorithm), and consider the full distribution over possible values of missing attributes when training the classifier [74, 66]. A third direction, and the focus of the present work, jointly learns the parameters of the imputation model and the classifier [47, 22, 71].

This joint learning of the imputation model and classifier has been shown to be advantageous for a variety of reasons [22]. First, sequential optimization (learning the imputation model first, followed by the classifier) is prone to compounding errors: mistakes made in learning the imputation model parameters are propagated into the training of the classifier. Joint optimization can avoid this issue by taking into account the effect of the imputation model on classification performance when estimating the imputation model parameters. Second, different choices of imputation models may be better suited to different classifiers, as some classifiers may be better at handling certain kinds of errors than others. Joint optimization addresses these issues by considering the effect of imputation model on classification performance during the parameter learning.

While several methods have been developed to jointly learn imputation model and classifier parameters, they have critical limitations that prevent applicability to real-world datasets. These methods either make assumptions about the data missingness that generally do not hold for medical data, or they cannot be used when there are missing values in the test data, which is the main use case for classification with missing values in most hospital settings.

In this chapter, we address these limitations and propose a method for joint optimization of the imputation model and classifier that is applicable to many medical datasets. Our proposed method is an optimization over both imputation model and classifier parameters, which uses the effect of the imputation model on classification loss to guide the solution towards imputation parameters that achieve better classification performance. Unlike existing methods for joint optimization, the proposed method makes no assumptions about the missingness mechanism of the data, making it appropriate for data that is not missing at random (NMAR). It can also be used when both training/testing data have missing values. This method can be used with a variety of choices of imputation model or classification loss functions and allows the joint optimization of probabilistic imputation models with discriminative classifiers.

The presence of missing data is particularly troublesome when stratifying trauma patients for adverse outcomes. Trauma is the leading cause of death in the United States in people under 44 years of age, and costs approximately $80 billion in medical treatment a year [26]. Trauma patients require urgent care and are often in very poor condition at their arrival to the hospital. As a result, it is often impossible to collect some clinical measurements at the time of admission. This missingness cannot be reduced through more staff or equipment as is the case in many other domains. This prevalence of missing values is a significant obstacle to the application of machine learning in trauma care, as the vast majority of methods do not work on data with missing values. We evaluate the proposed method for the prediction of adverse outcomes in trauma settings, and find that it significantly improves the accuracy of risk estimates relative to standard imputation approaches.

4.2 Background

4.2.1 Missingness

The canonical definition of missingness mechanisms comes from Little and Rubin [49]. Consider a dataset X, with all values present. The missingness matrix R is defined to be the same size as X, where

$$R_{ij} = \begin{cases} 0 & \text{if } X_{ij} \text{ is missing} \\ 1 & \text{if } X_{ij} \text{ is observed} \end{cases} \tag{4.1}$$

Define X^o as the portion of X that is observed (where $R = 1$), and X^m as the portion of X that is unobserved. The joint distribution of X and R can be modeled as

$$P(X, R|\theta, \phi) = P(X|\theta)P(R|X, \phi) \tag{4.2}$$

where θ are parameters governing the data distribution, and ϕ are parameters governing the missingness R, which may depend on the values in the data X. The missing completely at random (MCAR) mechanism assumes that the missingness is independent of the data:

$$P(R|X, \phi) = P(R|\phi) \tag{4.3}$$

A weaker assumption, and one that subsumes MCAR, is the missing at random (MAR) mechanism, which allows the missingness to depend on the observed, but not the unobserved values of X

$$P(R|X, \phi) = P(R|X^o, \phi) \tag{4.4}$$

When even this this MAR condition does not hold, the data is said to be not missing at random (NMAR). In the NMAR setting, the distribution of a missing variable conditioned on the observed variables may differ from the conditional distribution of that variable had it been observed. This makes it impossible to estimate the distribution from the data without information about the missingness mechanism. For example, if older respondents to a patient survey are less likely to report their age, the average of the missing age values will be higher than that of the observed age values. In the NMAR setting it is difficult to learn good estimates of the imputation model parameters θ, as the distributions of the data X and missingness R become coupled, requiring an explicit model of the missingness process.

4.2.2 Missing Value Imputation

Many methods exist for imputing missing values [29]. These range from simple approaches like filling missing values in with the mean of the observed data, to more complex approaches based on weighted averaging of the k-nearest neighbors or random forests.

The most commonly used approaches for imputation rely on statistical models of the data and use data distribution parameters to either fill in missing values with point estimates (e.g., their expected value conditioned on the observed values), or to generate many samples that approximate the distribution over possible values (multiple imputation). The EM algorithm is the most frequently used method for learning these data distribution parameters in the presence of missing values. Given a flexible class of models, such as the commonly used Gaussian mixture model, EM can easily learn models of the data distribution [73, 30]. Once the parameters of the distribution have been learned, these models can be used to impute the missing values, or reason about their uncertainty when estimating statistics from the data. Most applications of EM to imputation take advantage of the MAR assumption, seeking to optimize the full data likelihood over choices of model parameters by assuming the missing and observed data share the same distribution. It is clear in the case of NMAR that the model parameters that maximize the observed data likelihood may differ, possibly substantially, from the true parameters. While it is possible to use EM without making the MAR assumption, it requires defining a model for the generation of missingness in the data, which is challenging as the mechanism is often prohibitively complex or unknown.

One straightforward approach to handling NMAR data is the inclusion of missingness indicator variables: constructing R from the data and adding this missingness matrix to the set of predictive variables. This can improve performance when missingness is related to the class labels, however it does not help find a better value for θ, and the addition of missingness indicators cannot in general correct the noise induced by an erroneous imputation model.

In this work we focus on single imputations, however multiple imputation methods are also used in practice [67, 13]. However, as with EM-based single imputations, avoiding the MAR assumption with these methods usually requires making explicit assumptions about the missingness mechanism behind the data, and they ignore the choice of classifier or classification performance when learning the model parameters.

4.2.3 Joint Learning of Imputation Model and Classifier

In the standard sequential approach to learning imputation model and classifier, the imputation parameters θ are first learned using EM, and a classifier is then trained given the learned values of θ. This process may involve using point estimates of the missing values (their expected values under θ conditioned on the observed values), or using a classifier designed to account for the conditional distribution over missing values [74, 66].

In contrast, a joint learning of the imputation model and classifier can help find better choices of θ. Inaccuracies in imputation add noise to the classification task reducing accuracy. By accounting for classification loss when learning the imputation model during joint optimization, we can therefore avoid the pitfalls of a sequential approach (e.g., compounding errors, ignorance of classifier choice).

Liao et al. developed a graphical model which incorporated both the data distribution and the classification task [47]. When estimating this unified model's parameters using EM, the imputation model and classifier are learned together. This method uses a monolithic probabilistic model, and cannot be applied with different classification loss functions, like hinge loss or different regularization terms. Because this unified probabilistic model does not explicitly model the missingness mechanism in the data, it depends upon the MAR assumption, limiting the method's applicability to NMAR data.

Dick et al. proposed a method that learns an exact imputation of the missing values in the training data using an objective that takes the classification loss into account [22]. This method allows for flexibility in choice of classification loss function, and does not make the MAR assumption. Unfortunately, it does not learn a parametric imputation model, but rather an exact imputation for the training data. As a result, the method is unusable when there are missing values in the testing data. Complete test data is frequently unavailable in many medical scenarios, precluding use of the method in such cases.

We note that when the MAR condition is violated, maximizing the observed data likelihood may not help (and could even hurt) the accuracy of the imputed values. In the context of classification however, we have an additional metric (loss function) that can aid in learning good imputation models, and improve our handling of NMAR data by relying less on the observed data likelihood. A more accurate imputation model introduces less noise to the classification process, and allows for better predictions. For this reason, the effect of the imputation model on classification loss can help guide our choice of imputation parameters, even when we have no information about the missing data distribution.

4.3 Methods

We propose a method that addresses these limitations and extends the applicability of joint optimization of imputation model and classifer to additional kinds of data, and in particular many real-world medical datasets. The proposed method is an optimization problem over both imputation model and classifier parameters, which uses the effect of the imputation model on classification loss to guide the solution towards imputation model parameters that achieve better classification performance. Unlike existing approaches to joint optimization, the method makes no assumptions about the missingness mechanism of the data, making it better suited to data that is not missing at random (NMAR), and can be used when both training/testing data have missing values. The method can be used with a variety of choices of imputation model or classification loss functions, and allows the joint optimization of probabilistic imputation models with discriminative classifiers. The proposed joint optimization problem is presented in Equation 4.5. Let X denote the dataset, with X^o representing the observed portion of X. Y denotes the class labels, θ are the parameters of the imputation model, and w are the classifier parameters.

$$\underset{\theta,w}{\operatorname{argmax}}(1-\alpha)LL(X^o|\theta) - \alpha Loss(Y|X^o, \theta, w) \tag{4.5}$$

Equation 4.5 seeks to maximize a convex combination of the observed data log likelihood under the imputation model parameters θ (the first term), and the classification loss using both the classifier parameters w and the data imputation under θ conditioned on the observed data X^o (the second term). Unlike with a unified probabilistic model, learning the parameters in this combined objective function does not require making probabilistic assumptions about the missingness mechanism of the data, as in the work of Liao et al. [47], because the second term in the optimization is not probabilistic. However, in contrast to the work of Dick et al. [22] we assume the

Algorithm 1 Alternating optimization of Equation 4.5

```
**Input:** $X$, $Y$, $\alpha$, $K$
randomly initialize $\theta_0$ (with $K$ components)
$w_0$ = TrainClassifier($X^o$, $E[X^m|X^o, \theta_0]$, $Y$)
$i = 0$
**repeat**
  $i = i + 1$
  $\theta_i = \text{argmax}_\theta (1 - \alpha) LL(X^o|\theta) - \alpha Loss(Y|X^o, \theta, w_{i-1})$
  $w_i$ = TrainClassifier($X^o$, $E[X^m|X^o, \theta_i]$, $Y$)
**until** $\theta$ and $w$ have converged
**Output:** $\theta_i$, $w_i$
```

existence of a parametric imputation model, described by θ, which allows us to use the learned model parameters to generate imputations on unseen data with missing values.

The tuning parameter $\alpha \in [0, 1]$ controls the relative strength of the classification loss and observed data likelihood in learning the parameters. In the case where α is (nearly) equal to zero, the approach is equivalent to the traditional sequential optimization approach. The values of θ are determined entirely by the observed data likelihood, as in EM, and the classifier parameters w are optimized given that value of θ. When $\alpha = 1$, the classification loss alone guides the choice of parameters, effectively assuming that the observed data likelihood has no relevance. In this case, Equation 4.5 will try to find an imputation model that maximizes the separability of the data.

Decreasing the value of α can be thought of as a kind of regularization. Total reliance on classification loss to learn both the classifier and imputation model, which may have many parameters (particularly in high-dimensional data), is likely to lead to overfitting in smaller datasets. Reducing α encourages the imputation model to be closer to the observed data distribution. This can be loosely interpreted as treating the EM solution as a kind of prior on the missing data distribution, with α controlling the strength of that prior belief.

Equation 4.5 can be used with a variety of loss functions (e.g. log loss, hinge loss, or squared loss in the case of regression), as well as different data models (e.g. GMMs, multinomial mixture models). This framework allows for the combination of probabilistic imputation models with discriminative classifiers, and allows flexibility in the kind of regularization used for the classifier. In optimizing Equation 4.5, we take an alternating optimization approach, as shown in Algorithm 1. After generating intial estimates of the imputation model θ_0 and the classifier w_0, we alternate between optimizing the imputation model parameters θ conditioned on the current estimate of w, and optimizing the classifier parameters w using the imputation generated by the current imputation parameter estimate θ. The process repeats until the objective no longer improves.

In implementation, we consider the use of l_2 regularized log loss (logistic regression) for the classifier, and use GMMs for the data model. We implement the optimization of Equation 4.5 with respect to the GMM parameters θ given the classification weight vector w using gradient ascent. Estimation of w given θ was done using standard methods for training a logistic regression model, after filling in the missing values in X with their expected values. A validation set was used to select appropriate choices of α, the number of GMM components K, and the regularization parameter for the classifier. The optimization of Equation 4.5 is susceptible to local optima. As a result, we run the optimization with multiple random parameter initializations, and select the best model/classifier on a validation set.

4.4 Evaluation

We evaluated our proposed method on both synthetic and real-world data. We first present results on a synthetic dataset comparing sequential optimization using EM-based imputations with the proposal joint optimization algorithm on artificially generated missingness varying from MCAR to NMAR. This serves as an illustrative example of the method, and as an investigation into the relationship between α, the missingness mechanism, and the method's performance.

As our primary evaluation of the method, we compare sequential and joint optimization for the prediction of several adverse patient outcomes in a large national representative cohort of trauma patients. This evaluates the merits of the approach on real-world hospital data in several important prediction tasks.

4.4.1 Synthetic Data Evaluation

Synthetic data were used to compare the effect of missingness mechanism (MCAR vs. NMAR) on the method's performance, and particularly on the optimal choice of α. 1,000 data points were sampled from a 2-dimensional, 2-component mixture of Gaussians. The labels were generated by a perfect linear separator (see left panel in Figure 4.1).

Two different missingness mechanisms were used to generate the data. For the MCAR case, values were removed from the variable x_2 from all points with equal probability. To generate NMAR data, values were removed from the variable x_2 only in points generated by one of the components. Points generated by the other component had no missing values. As a result, the joint distribution of the variables x_1 and x_2 differed between the fully and partially observed data points.

To assess the effect of the amount of missing data on the method's performance, we varied the percentage of data points in which x_2 was missing. Once this percentage reaches 50% in the NMAR case, x_2 is entirely unobserved for points generated by the component with missing values, and the choice of imputation model becomes irrelevant. As a result, we limited the maximum percentage of points missing values for x_2 to 40%.

We compared the performance over a range of choices of α between 0 (equivalent to sequential optimization using EM) and 1 (fully loss-based approach), to see whether an objective based on classification loss could select a good imputation model and classifier. The two methods were compared over 20 random splits of the data into training (60%), validation (20%), and testing (20%) sets. Classification performance was measured using the area under the receiver operating characteristic curve (AUC), with the reported statistics reflecting the average across trials.

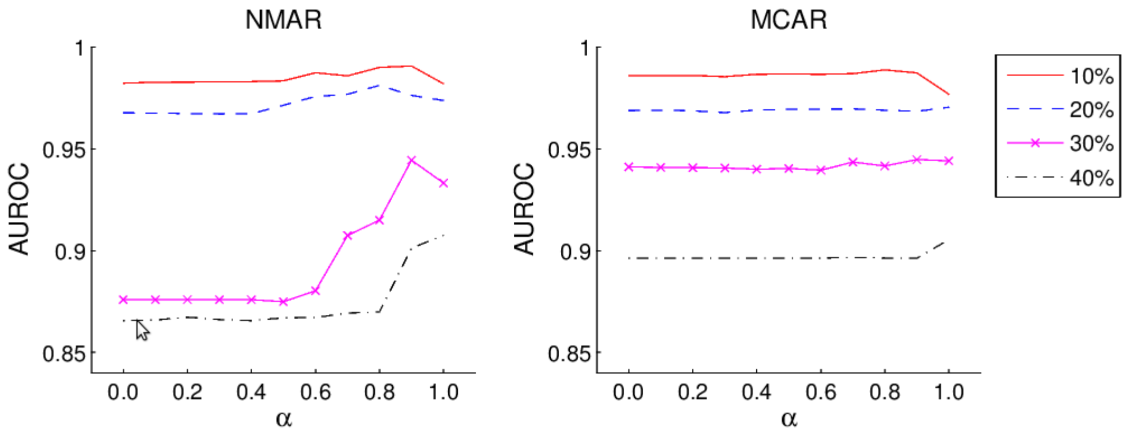

Figure 4.1: Classification performance using different choices of α on NMAR (left) and MCAR (right) data, for varying percentages of missingness.

Figure 4.1 shows the effect of the choice of α on AUC for both NMAR (left) and MCAR (right) missingness mechanisms (center and right panels). When the missingness was generated MCAR, there was a very minor improvement of α values greater than zero over the baseline sequential case ($\alpha = 0$). This is consistent with earlier work on joint optimization, which found small but statistically significant improvements in classification on MCAR data [22].

In the NMAR setting, α values greater than zero showed much larger improvements than in the MCAR setting. The improvement of joint optimization ($\alpha > 0$) increased with the amount of missing data, becoming particularly pronounced with 30% or more values missing. The loss term of Equation 4.5 is more sensitive to overfitting when fewer points have missing values, as it is determined by only a small number of data points. This justifies the use of smaller values of α with low amounts of missing data.

4.4.2 Prediction of Trauma Patient Outcomes

We investigated the performance of our approach at predicting several adverse outcomes in a national representative cohort of 162,821 trauma patients from the National Trauma Databank. The variables used included a variety of vital signs and scores collected at admission to the emergency department. The percentage of values missing for each attribute ranged from 3% to 40%. The missingness of variables in NTDB is typically due to an inability to collect these variables due to the patient's condition and is therefore highly likely to be NMAR. The outcomes of interest were whether a patient was admitted to the ICU (34%), whether they were put on a ventilator (17%), and whether they suffered in-hospital mortality (3%). For evaluation, the dataset was randomly divided into equally sized training, validation, and testing sets. The validation set was used to select α, the number of GMM components, and the classification regularization parameter.

For each clinical feature an indicator variable (0 or 1) was added to the training data to account for whether this variable was available or missing. Evaluation was conducted both with and without use of these missingness indicator variables to assess whether the proposed method provides complementary improvements for NMAR data beyond simply knowing that there was no opportunity to collect certain measurements.

Classification performance was measured using the area under the receiver operating characteristic curve (AUC). Additionally, the category-free net reclassification improvement (NRI) and integrated discrimination improvement (IDI) were used to assess the improvement of using joint optimization instead of sequential optimization. Details on these evaluation metrics are included in Appendix B.

Table 4.1 shows the AUC of various approaches on several patient outcomes on the NTDB dataset. The proposed joint optimization method achieved significantly higher AUC values than sequential optimization with mean imputation and EM imputation

	No Indicators		
	Seq. (Mean)	Seq. (EM)	Joint
ICU	0.676	0.681 (*ref.*)	**0.691** **(*0.004*)**
Ventilator	0.757	0.783 (*ref.*)	**0.794** **(*0.007*)**
Mortality	0.833	0.831 (*ref.*)	0.835 (*0.303*)
	Indicators		
	Seq. (Mean)	Seq. (EM)	Joint
ICU	0.684	0.685 (*ref.*)	**0.696** **(*0.002*)**
Ventilator	0.794	0.793 (*ref.*)	**0.798** **(*0.094*)**
Mortality	0.837	0.836 (*ref.*)	0.836 (*0.513*)

Table 4.1: AUC values for three outcomes on the NTDB dataset when using the sequential method with mean imputation [Seq. (Mean)], with EM imputation [Seq. (EM)], and the joint optimization method. P-values corresponding to the improvement of joint optimization over sequential are included in parentheses, with the baseline marked with (*ref.*). Results are shown with and without the use of missingness indicator variables. Bolded results indicate significant improvement at the 0.1 level.

	IDI (p)	
	No Indicators	Indicators
ICU	0.007 (*<0.001*)	0.008 (*<0.001*)
Ventilator	0.014 (*<0.001*)	0.005 (*<0.001*)
Mortality	0.001 (*<0.001*)	0.000 (*0.508*)
	NRI (p)	
	No Indicators	Indicators
ICU	0.120 (*<0.001*)	0.245 (*<0.001*)
Ventilator	0.574 (*<0.001*)	0.397 (*<0.001*)
Mortality	0.559 (*<0.001*)	0.268 (*<0.001*)

Table 4.2: Integrated discrimination improvement (IDI) scores (top) and net reclassification improvement (NRI) scores (bottom) with associated p values when assessing the change from sequential to joint optimization. Positive IDI/NRI values indicate an improvement in the accuracy of assigned probabilities by using joint instead of sequential optimization.

in predicting ICU admission and ventilator use. This improvement was consistent even after adding missingness indicator variables to the model. No significant difference was found between any of the methods in predicting in-hospital mortality. The presence of an improvement when using missingness indicator variables confirms that the missingness in the NTDB dataset is likely to be NMAR.

Table 4.2 shows the IDI and NRI of using the proposed method over the EM-based sequential method. The use of joint optimization led to significant improvements in IDI for all three outcomes of interest when missingness indicators were not used. When including missingness indicators, joint optimization had a statistically significant IDI for prediction of ICU admission and ventilator use, although not for mortality. Joint optimization achieved significant NRI values for all outcomes, regardless of whether missingness indicators were included in the models or not. These results indicate that the use of our method resulted in more accurate risk probability estimates than with the standard approach to handling missingness.

Joint optimization converged in fewer than 5 iterations (about 15 minutes to train on 50,000 examples using a 4-core Intel Xeon processor), taking approximately 20 times as long as the EM-based sequential method. We emphasize that the joint optimization training can be done offline with online evaluation of new patients taking the same amount of time as with standard methods.

4.5 Discussion

In this paper we present a general optimization problem that can jointly address imputation and classification. In contrast to existing methods for joint optimization, our method has several properties that make it applicable to real-world datasets: the proposed approach does not assume the data is MAR, is applicable when both training and testing data have missing values, and can use a variety of imputation models and classification loss functions.

Our method is motivated by the problem of evaluating trauma patients, for whom clinical variables frequently go uncollected due to the severity and urgency of their condition. When evaluated in a large and representative population of patients undergoing trauma surgery, our proposed approach achieved statistically significant improvements over standard sequential optimization in terms of the AUC, IDI, and NRI statistics, even with the addition of missingness indicator variables. While small in magnitude, even minor improvements in individual risk estimates could have a substantial effect on patient care and quality and outcomes initiatives (potentially affecting tens of thousands of patients per year).

Our evaluation showed greater benefits when using joint optimization on NMAR data in comparison to the commonly studied MCAR data on datasets with artificially generated missingness. These results confirm the intuition that as the distribution of the missing data differs from that of the observed, the MAR assumption harms the classification performance, and that the inclusion of classification loss can better reflect the utility of a choice of imputation model parameters.

The method was evaluated using single imputation for estimation of missing values, however future work could extend it to work with multiple imputations. The approach could also be used for regression, for example, by using squared error as the loss function in the joint optimization.

CHAPTER V

Conclusion

In this thesis we proposed several approaches that resulted in significant improvements to risk stratification across several clinical domains. The work presented included methods for learning novel risk factors, as well as methods for developing more accurate risk models. Collectively, this work demonstrates the potential for machine learning methods to advance clinical risk stratification, not only through incremental improvements in prediction accuracies, but by enabling substantial changes in the way that risk models are developed and applied in standard clinical practice.

5.1 Leveraging Physiological Recordings

Wearable devices that measure a variety of physiological properties are becoming more accurate and less expensive. These advances enable the collection of larger numbers of ambulatory recordings, and improve the feasibility of incorporating these recordings into standard clinical practice. However to translate these technological advances into real-world clinical impact we need computational methods that can not only make this large and complex data more interpretable to clinicians, but can use it to derive risk factors that improve upon risk models currently used in clinical practice. We proposed several approaches to learning structure in long-term physiological recordings. These approaches, while evaluated in the context of specific medical applications, are applicable across many kinds of physiological recordings and to a wide variety of clinical domains.

5.1.1 Learning Latent States

In chapter 2 we proposed the application of topic models to time series data to discover latent structure in the occurrence of large numbers of patterns. These models generate a condensed representation of a patient's underlying physiological state, by integrating information from many different properties of the recordings. The states learned by these models can serve as a visualization tool to improve the understanding of long recordings, giving clinicians a depiction of how the patient's condition varies over time, and what kinds of structure are present in a patient's recording. The topics, which are distributions over patterns, also provide some insight into patterns that may have an underlying physiological relationship with one another. When evaluated as a tool for visualization of sleep recordings, models of night-long EEG data identified richer structure than traditional sleep analysis methods. The use of these topics could help characterize sleep disorders that have been previously difficult to understand with the standard staging system.

The use of these topics for visualization has potential applications in visualizing many kinds of complex long-term recordings, including ECG, EEG, and the multi-modal recordings collected in intensive care units. Selecting the regions of recordings associated with particular topics could allow fast interpretation of the data by clinicians, and could help automatically identify unusual regions in long recordings to be presented to clinicians for manual inspection.

The topics learned by these models are also useful as risk factors for the prediction of adverse outcomes. When evaluated for the prediction of cardiovascular death from long-term heart rate, these models were more discriminative than previously proposed heart rate risk factors.

These topics could also be used to find relationships between the topics present in a patient's recording and the effectiveness of specific treatments, which could help identify subsets of patients who benefit most from particular medications or procedures.

The applications discussed used only a single physiological measurement in the model, either heart rate or EEG, and treated all possible fixed-length sequences of discretized features as patterns. The generality of the approach means that it is possible to explore broader definitions of patterns, and to incorporate multiple kinds of patterns in a single model. This could include the incorporation of multiple physiological parameters, like both heart rate and morphologic features of the ECG for cardiac risk stratification, or electrooculogram or ECG recordings with the EEG for sleep analysis. Patterns could be defined at multiple time scales, spanning multiple physiological parameters, could include specially defined patterns like those used to compute heart rate turbulence or be in both the time and frequency domains. Integration of a wider range of patterns could identify novel relationships that shed light on a patient's overall physiological state.

5.1.2 Learning Risk Factors with CNNs

Chapter 3 investigated the use of more powerful methods for learning risk markers from physiological recordings, through the application of convolutional neural networks. When evaluated for the prediction of cardiac death from long-term heart rate, the CNN risk factor was more discriminative than previous heart rate risk markers, including the topic models presented in chapter 2. Not only was the CNN risk factor the most useful heart rate risk factor, but it also achieved better discrimination than standard risk scores, and significantly improved prediction of cardiac death when included alongside these standard risk factors. This demonstrates the depth of untapped information about patient risk present in long-term physiology, and illustrates how more powerful approaches to learnining risk factors can improve our ability to leverage this data to risk stratify patients.

A prospective validation of the CNN risk score on new data is needed to confirm the discriminative power of the approach. Exploration of different loss functions, for example maximizing performance at different levels of precision or recall, could lead to more effective risk models for particular problems, such as deciding which patients should be prescribed implantable cardiac defibrillators. Learning these networks using a larger collection of patient data may enable the learning of deeper networks with more complex structure, resulting in even more useful risk factors.

An area for future investigation is the learning of features that synthesize information from both the heart rate and the beat to beat morphology of the ECG. Experiments discussed in chapter 3 did not find a benefit from combining pairwise morphologic distances with heart rate, however morphologic distance features are highly susceptible to noise. Is it possible that investigation of alternative measures of morphology (e.g. QT-intervals) as additional inputs to the CNN could lead to the discovery of more discriminative features by identifying high-risk structure that spans both heart rate and ECG morphology, which is not done in existing ECG-derived risk markers. When combined with previously proposed heart risk factors and several standard clinical risk scores, both the topics score and the CNN score were the only heart rate measures selected in the final model.

The use of CNNs has the potential to improve risk stratification using a variety of physiological measurements other than heart rate. Blood pressure and blood glucose, multimodal recordings from the ICU, metabolite levels, and EEG are all physiological measurements in which these networks can be used to predict patient outcomes of great clinical importance.

A critical challenge in the use of these models is finding ways to make their parameters and the structure learned more interpretable to clinicians. Expert understanding of what makes the model's predictions more effective than existing methods is not only a critical step in validating these models, but can also provide a greater understanding of how different patterns in the data relate to a patient's physiology and their risk of adverse outcomes.

5.2 More Patient Data Means More Missing Data

As electronic health records become more fully integrated into clinical practice, there is more data is available for the development of risk models, and it is easier to incorporate these risk models into clinical decision making. The increase in the amount of data collected is not just in the number of patients, but in the aggregation of many different kinds of information about a patient, including their medical history, medications, imaging, test results, notes from doctors and nurses, and diagnoses, and treatments. However the more variables that can be recorded for a patient, the more missing values will be present in the data: not all patients' medical history will be available, medication information may be incomplete, and tests and imaging may not have been ordered for a variety of reasons.

The ability to handle missing data doesn't just affect the application of risk models, but it can have a profound effect on the way that these models are developed as we collect more kinds of information about patients. Normally when models are developed, the risk factors considered for inclusion are those which are collected as part of standard practice and are available for most, if not all patients. However there are often many additional variables, with independent predictive value, that are not considered for inclusion because they are unavailable for many or even most patients. The ability to develop risk models that consider a wide range of risk factors without requiring that they all be observed would enable the inclusion of as much information as possible when estimating risk for a given patient. This ability would result in more accurate predictions for large numbers of patients.

Using models that are robust to missing values in the data would result in a fundamental change in the way that risk scores are developed and applied, by taking full advantage of the advent of electronic health records. Chapter 4 described a novel method for jointly learning imputation and classification models to improve the accuracy of risk models in the presence of missing observations. The approach significantly improved the accuracy of adverse outcome predictions in a large national registry of trauma patients, demonstrating how more careful handling of variables that are unavailable for many patients can result in more accurate risk estimates.

While we focused our evaluation of the approach in the context of trauma patients, this approach has broad applicability to learning risk models throughout medicine. The method could easily be extended for use in the regression of continuous outcomes, such as a patient's length of stay in the hospital or the ICU, or to predict the number of days until readmission to the hospital. In addition, this method has value not only for learning clinical risk models, but in the broader context of classification or regression in the presence of missing data.

APPENDIX

Clinical Risk Score Evaluation Metrics

In many clinical settings it is important to evaluate and compare risk models according to several criteria.

- *Discrimination* - How well can a risk factor distinguish positive examples from negative ones? How can we assess whether a proposed model is significantly better at distinguishing between positive and negative examples than a baseline model?

- *Calibration* - How accurately does a model's predicted probability of events match the actual likelihood of an event occurring? Accurate probabilities are key for clinical decision making, as the best choice of treatment for a patient with a 5% risk of events may differ from the best choice for a patient with a 25% risk of events. A model may be able to perfectly discriminate between events and non-events but be poorly calibrated (e.g. the model's predictions scale exponentially and not linearly with patient risk).

- *Survival analysis* - How well can a risk factor estimate a patients risk of events over time? Most clinical outcomes are not truly binary labels, as adverse events occur at different times for different patients. Several issues complicate the analysis of estimating patient risk with time-related outcomes, which require special methods to evaluate models in this setting. Clinical trials have limited durations, and when a study ends or a patient drops out of the study early there is no way to know if or when a patient will suffer the event of interest. There may be competing risks that affect the outcome of interest: a patient in a study investigating cardiovascular death may die from unrelated causes.

Each of these dimensions has one or more metrics that are commonly used in the clinical literature to assess the utility of risk factors.

A.1 The Area under the ROC Curve

Measures like error rate and accuracy are less useful in many clinical problems for several reasons. First, events of interest are often rare, which makes accuracy or error rate less meaningful (as always predicting no events is likely to be near optimal), and reduces the magnitude of improvements seen. Second, these metrics require defining a prediction threshold. In many cases, it is not clear what the desired precision or recall of a risk factor are, or they may depend on the specifics of a particular application. It is desirable to have a measure that assesses a model's discriminative power across a range of possible thresholds.

Area under the receiver operating characteristic curve (AUROC or AUC) is the commonly used alternative to accuracy in these settings. The ROC curve consists of the sensitivity and 1-specificity of the measure when evaluated at all possible thresholds. The AUC ranges from 0 to 1 (practically speaking from 0.5 to 1), with 0.5 reflecting totally random, uninformative predictions, and 1 reflecting the ability of the score to perfectly separate positive and negative examples. The AUC corresponds to the probability that a randomly selected positive example will have a higher value for the score than a randomly selected negative example.

The AUC is closely related to the Wilcoxon rank-sum test (also known as the

Mann-Whitney U test), which is a non-parametric test of whether the distribution of the positive class is stochastically greater than the distribution of the negative class (e.g. the median score of the positive class is greater than the median score of the negative class). This relationship was used by Delong et al. to devise a statistical test for the improvement of one scores AUC relative to a baseline scores AUC citedelong1988comparing. The AUC and the method of Delong et al. for comparison of AUCs are used throughout this thesis.

While a useful tool for measuring the discriminative ability of predictive models, the AUC can be challenging to use when comparing models. The standard way to compare models before and after the addition of a new feature is to look at the difference of the model AUC before and after the addition. However when the baseline model performs well, the addition of an independent and highly predictive variable to the model will likely result in a small increase to the AUC, much smaller than if that independent variable were added to a model with lower initial performance (e.g. AUC around 0.6). It takes a very large effect size for a new variable to generate a substantial increase to the AUC when baseline discrimination is good. The AUC and the difference in AUCs between two models is also difficult to relate to real-world effects. It is not clear what an improvement in AUC from 0.80 to 0.81 means in terms of the effect on actual predictions.

A.2 Alternatives for Comparing Discrimination between Two Models

To address these concerns, Pencina et al. proposed two new measures to assess the incremental improvement of adding a new risk factor to a model with standard risk factors, called the integrated discrimination improvement (IDI) and the net reclassification improvement (NRI). Unlike AUC, which considers only the ordering of data points, these metrics use the model's predicted probabilities, the accuracy of which is of great practical importance in the medical domain. At a high level, these metrics consider improvements in predicted probabilities: what percentage of patients with events are considered higher risk by the new model, and what percentage of patients without events are considered lower risk by the new model? The key difference between IDI and NRI is that NRI considers only the *direction* of an improvement, while IDI considers the *magnitude* of the improvement.

A.2.1 Net Reclassification Improvement

There are two variants of NRI, the standard version, and category-free or continuous NRI. The standard NRI score does not consider raw predicted probabilities, but instead uses risk groups. For example, if two models are used to divide patients into high, medium, and low risk groups, the superior model will reclassify patients who suffer events into higher risk categories than the inferior model, and reclassify patients without events into lower risk categories. Category-free NRI applies this same concept to non-dichotomized prediction, instead looking at the direction of change in raw probabilities: if a patient who suffers an event has a higher probability of events in the proposed model, their risk was moved in the right direction.

More concretely, category-free NRI measures the proportion of patients x whose estimated risk probabilities under a new model, $\hat{p}_{new}(x)$, are more accurate than estimates from an old model, $\hat{p}_{old}(x)$. The direction of change in estimated risk, $v(x_i)$, for a patient x_i is defined as follows:

$$v(x_i) = sign(\hat{p}_{new}(x_i) - \hat{p}_{old}(x_i)) \tag{A.1}$$

Instead of using the direction of change in predicted probabilities, standard NRI defines $v(x_i)$ as 1 if data point i moved to a higher risk group, and -1 if data point i moved to a lower risk group.

NRI combines the percentage of patients with events ($y_i = 1$) whose risk scores increase under the new model, with the percentage of patients without events ($y_i = 0$) whose risk scores decrease:

$$NRI = \frac{\sum_{i,\ y_i=1} v(x_i)}{\sum_k [y_k = 1]} - \frac{\sum_{j,\ y_j=0} v(x_j)}{\sum_k [y_k = 0]} \tag{A.2}$$

A.2.2 Integrated Discrimination Improvement

IDI instead focuses on the magnitude of improvements in the estimated risk probabilities:

$$IDI = \frac{\sum_{i,\ y_i=1} \hat{p}_{new}(x_i) - \hat{p}_{old}(x_i)}{\sum_k [y_k = 1]} - \frac{\sum_{i,\ y_i=0} \hat{p}_{new}(x_i) - \hat{p}_{old}(x_i)}{\sum_k [y_k = 0]} \tag{A.3}$$

A.2.3 Practical Considerations

The use of predicted probabilities in these metrics means that it is essential that both models being compared are well calibrated. If models are not appropriately calibrated, this can lead to misleading IDI and category-free NRI values. IDI and category-free NRI values should be accompanied by a measure that assesses whether the models compared were well calibrated.

The magnitude of a change in the AUC when adding a variable to a model is highly dependent on the strength of the baseline model. IDI is less sensitive than the AUC in this respect, but is not entirely independent of the baseline strength. The magnitude of the IDI is also sensitive to the baseline event rate: improvements in the prediction of rare events, where predicted probabilities should be low in magnitude, will be much smaller than improvements in the prediction of more common events. The NRI is not sensitive to either the strength of the baseline model or to the rates of events.

A.3 Measuring Calibration

The most commonly used method to evaluate calibration is the Hosmer-Lemeshow χ^2 score. Designed as a goodness-of-fit test for logistic regression models, this score ranks data points based on their predicted risk, divides them into risk deciles, and compares observed and expected event rates in each group. The test evaluates whether observed and expected rates differ significantly across the groups. The Hosmer-Lemeshow test statistic is defined as follows:

$$H = \sum_{g=1}^{G} \frac{(O_g - E_g)^2}{N_g \pi_g (1 - \pi_g)} \tag{A.4}$$

where G is the number of risk groups used (generally 10), O_g is the observed number of events in group g, E_g is the expected number of events in group g, N_g is the number of patients in group g, and pi_g is the predicted rate of events in group g. This statistic follows a χ^2 distribution with $G - 2$ degrees of freedom.

The score does not handle very large datasets well, because the more samples that are present in each risk decile, the more closely the observed event rate has to match the expected event rate to avoid finding a significant difference. For a fixed model, as the size of the dataset used in evaluation increases, the Hosmer-Lemeshow χ^2 score becomes more significant, which suggests that the model is more poorly calibrated. For this reason, application of the score to very large datasets (e.g. 100,000s of patients) almost always suggests that even the best models are poorly calibrated, and so the test results is often not meaningful in these settings.

A.4 Survival Analysis

Timing is an intrinsic aspect of most events of interest. The amount of time until an event occurs is meaningful, as a patient who suffers an event one month into a study may have much higher risk than a patient who suffers an event one year into the study. The limits of follow-ups in clinical trials mean that a patient may suffer the event of interest after the study concludes, and treating this data point as a binary outcome would consider them as a negative example. Some patients also drop out trials early. There may be competing risks, for example if your outcome of interest is cardiovascular death, but a patient dies for other reasons. Defining a timed events problem as a strict classification problem may induce label noise due to variations in censoring times between patients, and may lose information in these distinctions.

A.4.1 Harrell's C-statistic

One approach to incorporating timing information when evaluating risk factors is the use of Harrells C-statistic, or concordance C. When no censoring information is included, Harrells C-statistic is equivalent to the AUC, evaluating the likelihood a random positive example has a higher score than a random negative example. When censoring is included, the C-statistic considers the likelihood a random positive example has a higher score than a random negative example, provided that the positive example had an event time earlier than the censoring time of the negative example (e.g. a patient suffering an event 3 months into the follow-up would not be compared to a patient who suffered no events but dropped out of the study after 2 months).

A.4.2 Hazard Functions

The area of survival analysis attempts to model the amount of time before events happen. This can be approached through estimation of the survival function, which estimates the likelihood that an event occurs after a given time (e.g. the complement of an event not occurring by a given time). A related measure is the hazard function, which measures the probability of an event occurring at time t, given that an event has not occurred before time t. The cumulative hazard function is the integral of the hazard function up to a time t, and represents the likelihood of a patient suffering an event by time t (e.g. if the cumulative hazard function at 1 month is 50

It is possible to relate variables of interest (i.e. risk factors) to their effects on the hazard function of a population. A common approach to doing so is to assume proportional hazards: that risk factors have a multiplicative effect, called the hazard ratio, on the hazard function independent of time. For example, if a binary variable indicating whether a patient is over the age of 65 has a hazard ratio of 2, it means that being over 65 doubles the chance of the event of interest occurring at all times. Under the proportional hazards assumption, it is possible to estimate the multiplicative effect of parameters on the hazard function without any knowledge about the hazard function itself. This approach is called the Cox proportional hazards model. The equation for the hazard function λ at time t under the Cox proportional hazards model is:

$$\lambda(t|X) = \lambda_0(t) \exp(\beta_1 X_1 + \ldots + \beta_p X_p) \tag{A.5}$$

where λ_0 is the underlying hazard function for the data (and does not need to be estimated), X is a matrix of p predictive variables, and $\beta_{1\ldots p}$ are the coefficients associated with each predictor which are estimated from the data.

Given a dataset with event and censoring times along with a set of predictive variables for each patient, the hazard ratios (a function of $beta$) can be estimated under the Cox model, and the significance of their relationship to the occurrence of the event can be evaluated. Hazard ratios and associated p values under Cox proportional hazards models are used to compare the strength of the associations of various risk scores with outcomes of interest in chapters 2 and 4.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] E.M. Antman, M. Cohen, P.J.L.M. Bernink, C.H. McCabe, T. Horacek, G. Papuchis, B. Mautner, R. Corbalan, D. Radley, and E. Braunwald. The timi risk score for unstable angina/non–st elevation mi. *JAMA*, 284(7):835, 2000.

[2] S.P. Baker, B. O'Neill, W. Haddon Jr., and W.B. Long. The injury severity score: a method for describing patients with multiple injuries and evaluating emergency care. *Journal of Trauma and Acute Care Surgery*, 14(3):187–196, 1974.

[3] G.H. Bardy, K.L. Lee, D.B. Mark, J.E. Poole, D.L. Packer, R. Boineau, M. Domanski, C. Troutman, J. Anderson, and G. Johnson. Amiodarone or an implantable cardioverter–defibrillator for congestive heart failure. *New England Journal of Medicine*, 352(3):225–237, 2005.

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

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

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

[7] Y. Bengio, A. Courville, and P. Vincent. Representation learning: a review and new perspectives. 2013.

[8] D. Blei and J. McAuliffe. Supervised topic models. In *Neural Information Processing Systems*, 2007.

[9] D.M. Blei, A.Y. Ng, and M.I. Jorda. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.

[10] W. Braunwald. *Unstable angina: a classification*. Springer, 1990.

[11] E. Buccelletti, E. Gilardi, E. Scaini, L. Galiuto, R. Persiani, A. Biondi, F. Basile, and N.G. Silveri. Heart rate variability and myocardial infarction: systematic literature review and metanalysis. *Eur Rev Med Pharmacol Sci*, 13(4):299–307, 2009.

[12] J. Buckelmüller, H.P. Landolt, H.H. Stassen, and P. Achermann. Trait-like individual differences in the human sleep electroencephalogram. *Neuroscience*, 138(1):351–356, 2006.

[13] S. Buuren and K. Groothuis-Oudshoorn. Mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3), 2011.

[14] A.E. Buxton. Not everyone with an ejection fraction 30% should receive an icd. *Circulation*, 111(19):2537–2549, 2005.

[15] C.P. Cannon, S. Husted, R.A. Harrington, B.M. Scirica, H. Emanuelsson, G. Peters, and R.F. Storey. Safety and tolerability and and initial efficacy of azd6140 and the first reversible oral adenosine diphosphate receptor antagonist and compared with clopidogrel and in patients with non–st-segment elevation acute coronary syndrome: primary results of the disperse-2 trial. *Journal of the American College of Cardiology*, 50(19):1844–1851, 2007.

[16] H.R. Champion, W.J. Sacco, A.J. Carnazzo, W. Copes, and W.J. Fouty. Trauma score. *Critical Care Medicine*, 9(9):672–676, 1981.

[17] G. Chechik, G. Heitz, G. Elidan, P. Abbeel, and D. Koller. Max-margin classification of incomplete data. *Advances in Neural Information Processing Systems*, 19:233, 2007.

[18] C.C. Chia and Z. Syed. Computationally generated cardiac biomarkers: Heart rate patterns to predict death following coronary attacks. In *SIAM International Conference on Data Mining*, 2011.

[19] M.J. Cohen, A.D. Grossman, D. Morabito, M.M. Knudson, A.J. Butte, and G.T. Manley. Identification of complex metabolic states in critically injured patients using bioinformatic cluster analysis. *Critical Care*, 14(1):R10, 2010.

[20] M. Daley, C.M. Morin, M. LeBlanc, J.P. Gregoire, J. Savard, and L. Baillargeon. Insomnia and its relationship to health-care utilization and work absenteeism and productivity and accidents. *Sleep Medicine*, 10(4):427–438, 2009.

[21] R.J. de Winter and J.G.P. Tijssen. Non-st-segment elevation myocardial infarction: Revascularization for everyone? *JACC: Cardiovascular Interventions*, 5(9):903–905, 2012.

[22] U. Dick, P. Haider, and T. Scheffer. Learning from incomplete data with infinite imputations. In *International Conference on Machine Learning*, pages 232–239. ACM, 2008.

[23] A. Van Esbroeck, C.C. Chia, and Z. Syed. Heart rate topic models. *AAAI*, 2012.

[24] A. Van Esbroeck and Z. Syed. Cardiovascular risk stratification with heart rate topics. *Computing in Cardiology*, 2012.

[25] A. Van Esbroeck and B. Westover. Data-driven modeling of sleep states from eeg. In *IEEE Engineering in Medicine and Biology Society*, pages 5090–5093. IEEE, 2012.

[26] E.A. Finkelstein, P.S. Corso, and T.R. Miller. *The incidence and economic burden of injuries in the United States*. Oxford University Press, 2006.

[27] A. Flexer, G. Gruber, and G. Dorffner. A reliable probabilistic sleep stager based on a single eeg signal. *Artificial Intelligence in Medicine*, 33(3):199–207, 2005.

[28] K.A. Fox, O.H. Dabbous, J. Robert, K.S. Pieper, K.A. Eagle, F. van de Werf, A. Avezum, S.G. Goodman, M.D. Flather, and F.A. Anderson. Prediction of risk of death and myocardial infarction in the six months after presentation with acute coronary syndrome: prospective multinational observational study (grace). *British Medical Journal*, 333(7578):1091, 2006.

[29] P.J. García-Laencina, J.L. Sancho-Gómez, and A.R. Figueiras-Vidal. Pattern classification with missing data: a review. *Neural Computing and Applications*, 19(2):263–282, 2010.

[30] Z. Ghahramani and M.I. Jordan. Supervised learning from incomplete data via an em approach. In *Advances in Neural Information Processing Systems*, 1994.

[31] A.S. Go, D. Mozaffarian, V.L. Roger, E.J. Benjamin, J.D. Berry, M.J. Blaha, S. Dai, E.S. Ford, C.S. Fox, and S. Franco. Heart disease and stroke statistics 2014 update: a report from the american heart association. *Circulation*, 129(3):e28, 2014.

[32] D. Grangier and I. Melvin. Feature set embedding for incomplete data. In *Advances in Neural Information Processing Systems*, 2010.

[33] P. Halász. K-complex and a reactive eeg graphoelement of nrem sleep: an old chap in a new garment. *Sleep Medicine Reviews*, 9(5):391–412, 2005.

[34] P.S. Hamilton and W.J. Tompkins. Quantitative investigation of qrs detection rules using the mit/bih arrhythmia database. *IEEE Transactions on Biomedical Engineering*, (12):1157–1165, 1986.

[35] O. Hasan, D.O. Meltzer, S.A. Shaykevich, C.M. Bell, P.J. Kaboli, A.D. Auerbach, T.B. Wetterneck, V.M. Arora, J. Zhang, and J.L. Schnipper. Hospital readmission in general medicine patients: a prediction model. *Journal of General Internal Medicine*, 25(3):211–219, 2010.

[36] P.A. Heidenreich, J.G. Trogdon, O.A. Khavjou, J. Butler, K. Dracup, M.D. Ezekowitz, E.A. Finkelstein, Y. Hong, S.C. Johnston, and A. Khera. Forecasting the future of cardiovascular disease in the united states a policy statement from the american heart association. *Circulation*, 123(8):933–944, 2011.

[37] G. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002.

[38] J.L. Hossain and C.M. Shapiro. The prevalence and cost implications and and management of sleep disorders: an overview. *Sleep and Breathing*, 6(2):85–102, 2002.

[39] Y. Jia. Caffe: an open source convolutional architecture for fast feature embedding. http://caffe.berkeleyvision.org/, 2013.

[40] C.D. Johansen, R.H. Olsen, L.R. Pedersen, P. Kumarathurai, M.R. Mouridsen, Z. Binici, T. Intzilakis, L. Køber, and A. Sajadieh. Resting and night-time and and 24 h heart rate as markers of cardiovascular risk in middle-aged and elderly men and women with no apparent heart disease. *European Heart Journal*, 34(23):1732–1739, 2013.

[41] K.E. Joynt and A.K. Jha. Characteristics of hospitals receiving penalties under the hospital readmissions reduction program. *JAMA*, 309(4):342–343, 2013.

[42] K.E. Joynt, E.J. Orav, and A.K. Jha. Thirty-day readmission rates for medicare beneficiaries by race and site of care. *JAMA*, 305(7):675–681, 2011.

[43] Y. Kim, M. Kurachi, M. Horita, K. Matsuura, and Y. Kamikawa. Agreement in visual scoring of sleep stages among laboratories in japan. *Journal of Sleep Research*, 1(1):58–60, 1992.

[44] W.A. Knaus, E.A. Draper, D.P. Wagner, and J.E. Zimmerman. Apache ii: a severity of disease classification system. *Critical Care Medicine*, 13(10):818–829, 1985.

[45] H. Lee, R. Grosse, R. Ranganath, and A.Y. Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *International Conference on Machine Learning*, pages 609–616. ACM, 2009.

[46] H. Lee, P. Pham, Y. Largman, and A.Y. Ng. Unsupervised feature learning for audio classification using convolutional deep belief networks. In *Advances in Neural Information Processing Systems*, pages 1096–1104, 2009.

[47] X. Liao, H. Li, and L. Carin. Quadratically gated mixture of experts for incomplete data classification. In *International Conference on Machine Learning*, pages 553–560. ACM, 2007.

[48] J. Lin, E. Keogh, S. Lonardi, and B. Chiu. A symbolic representation of time series and with implications for streaming algorithms. In *Proceedings of the ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, pages 2–11. ACM, 2003.

[49] R.J.A. Little and D.B. Rubin. *Statistical analysis with missing data*, volume 4. Wiley New York, 1987.

[50] Y. Liu, Z. Syed, B.M. Scirica, D.A. Morrow, J.V. Guttag, and C.M. Stultz. Ecg morphological variability in beat space for risk stratification after acute coronary syndrome. *Journal of the American Heart Association*, 3(3):e000981, 2014.

[51] M. Malik. Heart rate variability. *Annals of Noninvasive Electrocardiology*, 1(2):151–181, 1996.

[52] P. Mirowski, D. Madhavan, Y. LeCun, and R. Kuzniecky. Classification of patterns of eeg synchronization for seizure prediction. *Clinical Neurophysiology*, 120(11):1927–1940, 2009.

[53] P.W. Mirowski, Y. LeCun, D. Madhavan, and R. Kuzniecky. Comparing svm and convolutional networks for epileptic seizure prediction from intracranial eeg. In *IEEE Workshop on Machine Learning for Signal Processing*, pages 244–249. IEEE, 2008.

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

[55] T. Penzel, K. Stephan, S. Kubicki, and W.M. Herrmann. Integrated sleep analysis and with emphasis on automatic methods. *Epilepsy Research*, 2:177, 1991.

[56] T.G. Pickering, W.B. White, and American Society of Hypertension Writing Group. When and how to use self (home) and ambulatory blood pressure monitoring. *Journal of the American Society of Hypertension*, 4(2):56–61, 2010.

[57] S.S. Rathore, J.M. Foody, Y. Wang, G.L. Smith, J. Herrin, F.A. Masoudi, P. Wolfe, E.P. Havranek, D.L. Ordin, and H.M. Krumholz. Race and quality of care and and outcomes of elderly patients hospitalized with heart failure. *JAMA*, 289(19):2517–2524, 2003.

[58] A. Rechtschaffen and A. Kales. *A manual of standardized terminology and techniques and scoring system for sleep stages of human subjects*. Number 204. US Department of Health and Education and Welfare and National Institute of Health, 1968.

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

[60] V.L. Roger, A.S. Go, D.M. Lloyd-Jones, R.J. Adams, J.D. Berry, T.M. Brown, M.R. Carnethon, S. Dai, G. de Simone, and E.S. Ford. Heart disease and stroke statistics 2011 update. a report from the american heart association. *Circulation*, 123(4):e18–e209, 2011.

[61] J. Santamaria and K. Chiappa. *Electroencephalography of drowsiness.* Demos Medical Publishing, 1987.

[62] S. Saria, D. Koller, and A. Penn. Learning individual and population level traits from clinical temporal data. In *Neural Information Processing Systems and Predictive Models in Personalized Medicine Workshop*, 2010.

[63] S. Saria, A.K. Rajani, J. Gould, D. Koller, and A.A. Penn. Integration of early physiological responses predicts later illness severity in preterm infants. *Science Translational Medicine*, 2(48):48ra65–48ra65, 2010.

[64] G. Schmidt, M. Malik, P. Barthel, R. Schneider, K. Ulm, L. Rolnitzky, A.J. Camm, J.T. Bigger, and A. Schömig. Heart-rate turbulence after ventricular premature beats as a predictor of mortality after acute myocardial infarction. *The Lancet*, 353(9162):1390–1396, 1999.

[65] A. Shoeb, H. Edwards, J. Connolly, B. Bourgeois, S.T. Treves, and J. Guttag. Patient-specific seizure onset detection. *Epilepsy & Behavior*, 5(4):483–498, 2004.

[66] A. Smola, S.V.N. Vishwanathan, and T. Hoffman. Kernel methods for missing variables. In *Advances in Neural Information Processing Systems*, 2005.

[67] J.A.C. Sterne, I.R. White, J.B. Carlin, M. Spratt, P. Royston, M.G. Kenward, A.M. Wood, and J.R. Carpenter. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *British Medical Journal*, 338, 2009.

[68] Z. Syed, C.M. Stultz, B.M. Scirica, and J.V. Guttag. Computationally generated cardiac biomarkers for risk stratification after acute coronary syndrome. *Science Translational Medicine*, 3(102):102ra95–102ra95, 2011.

[69] G. van der Heijden, R. Donders, T. Stijnen, and K. Moons. Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example. *Journal of Clinical Epidemiology*, 59(10):1102–1109, 2006.

[70] A.N. Vgontzas, D. Liao, E.O. Bixler, G.P. Chrousos, and A. Vela-Bueno. Insomnia with objective short sleep duration is associated with a high risk for hypertension. *Sleep*, 32(4):491, 2009.

[71] C. Wang, X. Liao, L. Carin, and D.B. Dunson. Classification with incomplete data using dirichlet process priors. *The Journal of Machine Learning Research*, 11:3269–3311, 2010.

[72] W.B. White, L. Wolfson, D.B. Wakefield, C.B. Hall, P. Campbell, N. Moscufo, J. Schmidt, R.F. Kaplan, G. Pearlson, and C.R.G. Guttmann. Average daily blood pressure and not office blood pressure and is associated with progression of cerebrovascular disease and cognitive decline in older people. *Circulation*, 124(21):2312–2319, 2011.

[73] D. Williams, X. Liao, Y. Xue, and L. Carin. Incomplete-data classification using logistic regression. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 972–979. ACM, 2005.

[74] D. Williams, X. Liao, Y. Xue, L. Carin, and B. Krishnapuram. On classification with incomplete data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3):427–436, 2007.

[75] P.W.F. Wilson, R.B. DAgostino, D. Levy, A.M. Belanger, H. Silbershatz, and W.B. Kannel. Prediction of coronary heart disease using risk factor categories. *Circulation*, 97(18):1837–1847, 1998.

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