Applying Domain Knowledge to Clinical Predictive Models

by

Yun Liu

B.S., Johns Hopkins University (2010)

Submitted to the Department of Health Sciences and Technology in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Medical Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2016

© Massachusetts Institute of Technology 2016. All rights reserved.

Author .................................................................
Department of Health Sciences and Technology
May 16, 2016

Certified by ..............................................................
Collin M. Stultz, MD, PhD
Professor
Thesis Supervisor

Certified by ..............................................................
John V. Guttag, PhD
Dugald C. Jackson Professor
Thesis Supervisor

Accepted by ..............................................................
Emery N. Brown, MD, PhD
Director, Harvard-MIT Program in Health Sciences and Technology/Professor of Computational Neuroscience and Health Sciences and Technology
Applying Domain Knowledge to Clinical Predictive Models

by

Yun Liu

Submitted to the Department of Health Sciences and Technology on May 16, 2016, in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Medical Engineering

Abstract

Clinical predictive models are useful in predicting a patient’s risk of developing adverse outcomes and in guiding patient therapy. In this thesis, we explored two different ways to apply domain knowledge to improve clinical predictive models.

We first applied knowledge about the heart to engineer better frequency-domain features from electrocardiograms (ECG). The standard frequency domain (in Hz) quantifies events that repeat with respect to time. However, this may be misleading because patients have different heart rates. We hypothesized that quantifying frequency with respect to heartbeats may adjust for these heart rate differences. We applied this beat-frequency to improve two existing ECG predictive models, one based on ECG morphology, and the other based on instantaneous heart rate. We then used machine learning to find predictive frequency bands. When evaluated on thousands of patients after an acute coronary syndrome, our method significantly improved prediction performance (e.g., area under curve, AUC, from 0.70 to 0.75). In addition, the same bands were found to be predictive in different patients for beat-frequency, but not for the standard frequency domain.

Next, we developed a method to transfer knowledge from published biomedical articles to improve predictive models when training data are scarce. We used this knowledge to estimate the relevance of features to a given outcome, and used these estimates to improve feature selection. We applied our method to predict the onset of several cardiovascular diseases, using training data that contained only 50 adverse outcomes. Relative to a standard approach (which does not transfer knowledge), our method significantly improved the AUC from 0.66 to 0.70. In addition, our method selected 60% fewer features, improving interpretability of the model by experts, which is a key requirement for models to see real-world use.

Thesis Supervisor: Collin M. Stultz, MD, PhD
Title: Professor

Thesis Supervisor: John V. Guttag, PhD
Title: Dugald C. Jackson Professor
Acknowledgments

They say that hindsight is 20/20. Even in hindsight, however, I cannot think of better co-advisers, than Collin and John. Their technical expertise, intuition, and understanding of the big picture have shaped all parts of my research. Their dedication to students in both technical and non-technical aspects is inspirational. I shall always cherish this experience and look forward to working with both Collin and John in the future.

I am also indebted to my thesis committee members: Thomas Heldt and Polina Golland. They have been instrumental in providing different perspectives both for the work and my future career. This work has also benefited greatly from (sometimes very lively) discussions with past and present members of both the Guttag group (Dorothy, Jenna, Anima, Garthee, Guha, Joel, Marzyeh, Jen, Amy, Tiam, Davis, and Maggie), and the Stultz group (Sarah, Linder, Diego, Virginia, Daniel, Orly, Thomas, Munishika, Paul, Gordon, Molly, and Kaustav). I am also thankful for the MEng and UROP students who have worked directly with me and assisted with exploring ideas that I otherwise would not have been able to: Mathura, Hussein, Jessica, Harlin, Maha, and Harini. Thanks also go to Arlene and Sheila and the HST office (Julie, Laurie, Traci, Joe, and Patty) for help in scheduling and navigating the administrative and academic requirements.

In my time at MIT, I’ve also had the privilege to collaborate with amazing researchers from various places. For the work on ECG signals, thanks go to Zeeshan Syed at the University of Michigan for developing some of the techniques that we built upon, and Ben Scirica at the TIMI group at Brigham and Women’s Hospital, for his clinical expertise. The work based on health insurance data was possible through collaborations and conversations with Robert Lu, Kun-Ta Chuang, and Chi-Hsuan Huang at the National Cheng Kung University in Taiwan, and Pete Szolovits at MIT. Ongoing work to collect and analyze ECG and physical activity data was possible through efforts of Charlie Sodini, Maggie Delano, Catherine Ricciardi, Sidney Primas, and Stas Lukashov at MIT.
Finally, I am deeply grateful for the support of my family and friends for their well wishes and support. Special thanks go to my partner Stephanie for her perspective and encouragement. I am also indebted to A*STAR, Singapore, for fellowship support both for my undergraduate and graduate training.
Contents

1 Introduction 19

1.1 Applying Medical Knowledge to Aid Feature Engineering . . . . . . . 21
   1.1.1 Motivation ........................................................................ 21
   1.1.2 Contributions ................................................................... 21

1.2 Transferring Medical Knowledge From Text .......................... 23
   1.2.1 Motivation ........................................................................ 23
   1.2.2 Contributions ................................................................... 24

1.3 Organization of Thesis .......................................................... 25

2 Background 27

2.1 Risk Stratification using Electrocardiograms (ECG) ............ 27
   2.1.1 The Heart and ECG ............................................................ 27
   2.1.2 Acute Coronary Syndrome (ACS) ...................................... 30
   2.1.3 Clinical Risk Metrics .......................................................... 31
   2.1.4 Long-term ECG Risk Metrics ............................................. 33
   2.1.5 Evaluation and Comparison of Risk Metrics ...................... 37

2.2 Risk Stratification using Health Insurance Records ........... 39
   2.2.1 Transfer Learning ............................................................... 39
   2.2.2 Distributed Representation of Words ................................. 41
   2.2.3 Word2vec ...................................................................... 42
   2.2.4 Word Vector Properties ................................................... 43
   2.2.5 Health Insurance Records ................................................ 43
### 3 Risk Stratification Using ECG: Morphological Variability in Beat-space

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1 Introduction</td>
<td>47</td>
</tr>
<tr>
<td>3.2 Methods</td>
<td>48</td>
</tr>
<tr>
<td>3.2.1 Dataset</td>
<td>48</td>
</tr>
<tr>
<td>3.2.2 Morphologic Variability in Beat-space (MVB)</td>
<td>50</td>
</tr>
<tr>
<td>3.2.3 Analysis of MVB in Specific ECG Segments</td>
<td>54</td>
</tr>
<tr>
<td>3.2.4 Correlation among Time and Beat-frequency Features</td>
<td>55</td>
</tr>
<tr>
<td>3.2.5 Comparison with Published Risk Metrics</td>
<td>55</td>
</tr>
<tr>
<td>3.2.6 Statistical Analysis</td>
<td>56</td>
</tr>
<tr>
<td>3.3 Results</td>
<td>57</td>
</tr>
<tr>
<td>3.3.1 Association between Risk Metrics and CVD in the validation cohort</td>
<td>57</td>
</tr>
<tr>
<td>3.3.2 Association between Risk Metrics and CVD in low-risk subgroups</td>
<td>57</td>
</tr>
<tr>
<td>3.3.3 Choosing an Optimal Threshold for MVB</td>
<td>59</td>
</tr>
<tr>
<td>3.3.4 Correlation of MVB with other Variables</td>
<td>60</td>
</tr>
<tr>
<td>3.3.5 Analysis of MVB in Specific ECG Segments</td>
<td>60</td>
</tr>
<tr>
<td>3.3.6 Correlation between Time and Beat-Frequency Bands</td>
<td>60</td>
</tr>
<tr>
<td>3.4 Discussion</td>
<td>62</td>
</tr>
<tr>
<td>3.4.1 Predictive Performance</td>
<td>62</td>
</tr>
<tr>
<td>3.4.2 Physiological Basis</td>
<td>63</td>
</tr>
<tr>
<td>3.5 Limitations</td>
<td>64</td>
</tr>
<tr>
<td>3.6 Future Work</td>
<td>65</td>
</tr>
<tr>
<td>3.7 Conclusion</td>
<td>66</td>
</tr>
</tbody>
</table>

### 4 Risk Stratification Using ECG: Heart Rate Variability

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1 Introduction</td>
<td>67</td>
</tr>
<tr>
<td>4.2 Methods</td>
<td>68</td>
</tr>
<tr>
<td>4.2.1 Data and Outcomes</td>
<td>68</td>
</tr>
<tr>
<td>4.2.2 Time-frequency LF/HF</td>
<td>69</td>
</tr>
</tbody>
</table>
4.2.3 Beat-frequency LF/HF ............................................. 70
4.2.4 Machine Learning .............................................. 70
4.2.5 Evaluation of Machine Learning Models ...................... 72
4.2.6 Comparison with Clinical Measures ......................... 73
4.2.7 Statistical analysis ............................................. 73
4.2.8 Correlation Between Time- and Beat-Frequency .......... 74

4.3 Results ................................................................. 75
4.3.1 LF/HF in Time- and Beat-frequency .......................... 75
4.3.2 Machine Learning (WHRV) in Time- and Beat-frequency .. 75
4.3.3 Correlation between Time- and Beat-frequency ............. 79

4.4 Discussion .............................................................. 80
4.4.1 Physiological Basis of Beat-Frequency Improvements ...... 81
4.4.2 Asymmetric Benefits of Machine Learning .................. 82

4.5 Limitations .............................................................. 83
4.6 Future Work ............................................................. 84
4.7 Conclusion ............................................................. 84

5 Risk Stratification Using Health Insurance Records: Transferring
Knowledge from Text ..................................................... 87

5.1 Introduction ............................................................. 87
5.1.1 Related Work ....................................................... 89

5.2 Data & Features ......................................................... 91
5.2.1 Data ................................................................. 91
5.2.2 Feature Transformation ......................................... 92
5.2.3 Features and Hierarchies ....................................... 94

5.3 Methods ................................................................. 95
5.3.1 Computing Estimated Feature-Relevance ..................... 95

5.4 Experiment Setup ....................................................... 101
5.4.1 Outcomes .......................................................... 101
5.4.2 Experimental Protocol .......................................... 102
5.5 Experiments and Results ............................................. 103
  5.5.1 Ranking of Features ........................................... 104
  5.5.2 Full Dataset .................................................. 104
  5.5.3 Downsampling Dataset with 50 Positive Examples .......... 104
  5.5.4 Downsampling Dataset with 25 Positive Examples .......... 106
5.6 Discussion .......................................................... 106
5.7 Limitations and Future Work ...................................... 109
5.8 Conclusion .......................................................... 110

6 Summary and Conclusion ............................................. 111
  6.1 Feature Engineering for the ECG ............................... 111
    6.1.1 Implications and Future Work ................................ 112
  6.2 Feature Selection Using Knowledge from Text .................. 113
    6.2.1 Implications and Future Work ................................ 114
List of Figures

2-1 The conduction system of the heart [3]. ........................................ 28
2-2 A normal ECG with labels on characteristic features. Figure adapted from [1]. ..................................................... 29
2-3 Heart rate in beats per min (bpm) for a health person. Figure edited from [4]. ..................................................... 30
2-4 Simplified decision tree for acute coronary syndrome. ......................... 30
2-5 A normal 12-lead ECG [2]. ..................................................... 33
2-6 Comparison between traditional machine learning and transfer learning [74]. ..................................................... 39
2-7 Word2vec architectures, figure from [62]. ........................................ 42
2-8 Example of hierarchy for ICD-9 diagnosis codes (left) and ATC medication codes (right). ........................................ 44
3-1 Because of instantaneous heart rate changes, cardiac events are periodic only with respect to heartbeats but not with respect to time. .... 50
3-2 Overview of Morphologic Variability (MV) and MV in Beat-space (MVB) computation. ..................................................... 51
3-3 Optimizing the diagnostic beat-frequency for maximum AUC in the derivation cohort. Our peak AUC is 0.73, at every 2 to 7 beats. The inset illustrates the Receiver Operating Characteristic (ROC) curve for this optimal diagnostic beat-frequency. ........................................ 54
3-4 Kaplan-Meier curves demonstrating risk stratification of two relatively lower risk subpopulations using the upper quartile value in each population (A: TRS ≤ 4; B: TRS ≤ 4 and BNP ≤ 80 pg/ml). Numbers of patients remaining in the study at each labeled time point are indicated below the respective labels.

3-5 Rate of cardiovascular death by quartiles of MVB. Lower-Risk-1 indicates TRS ≤ 4 and EF > 40%, Low-Risk-2 indicates BNP ≤ 80 pg/ml and TRS ≤ 4, and Lower-Risk-3 indicates BNP ≤ 80 pg/ml and EF > 40% and TRS ≤ 4. “missing” bars in the bottom two quartiles for the lower-risk-2 and lower-risk-3 populations indicate no deaths occurred in those subgroups.

3-6 Correlation between time-frequency bands (left) and beat-frequency bands (right).

4-1 Comparison of HRV metrics assessed in this study. In blue: time- and beat-frequency versions of LF/HF, the ratio of energy in two predefined frequency bands. In red: time- and beat-frequency versions of the energy in bands weighted using machine learning.

4-2 Boxplot of normalized weights of time- and beat-frequency machine learning models trained on the same 1,000 training splits. Quartiles are represented by the edges of lines, box, and central dot, while circles indicate outliers. The average standard deviations of the weights are 0.057 in beat-frequency and 0.153 in time-frequency.

4-3 Positive and negative predictive values of WHRV using different high risk thresholds.

4-4 Location of high frequency band in exercising subjects. Time-frequency data are from [76]: beat-frequencies are estimated by dividing the time-frequency value by the heart rate in Hz. Error bars represent standard error.
Overview of transferring knowledge from text to predict outcomes using a structured dataset.

Schematic of hill transform that interpolates between log-like shapes and sigmoid shapes.

Overview of computing and using the estimated relevance of each feature. The rescaled features in the final step are then used as input to a L1-regularized logistic regression model.

Word2vec maps words to lower dimensional representation (e.g., from a sparse binary vector of length 4 million to a dense vector of length 200).

Histograms of feature relevances computed using various means: arithmetic mean, power means with exponents 10 and 100, and maximum values.

Schematic of patient timeline used to extract features and define outcomes. \( k \) represents the number of occurrences of the appropriate billing codes (Table 5.2) in the five year period.

AUC and number of selected features using 50 positive examples using the standard approach compared with our proposed rescaling approach. Outcome and gender are labeled on the x-axis, each spanning two predictive tasks: age groups 1 and 2. Triangles indicate a significantly higher (pointed up) or lower (pointed down) AUC or number of selected features for our proposed method. Error bars indicate standard error.

AUC and number of selected features in the full and downsampled datasets averaged across all prediction tasks.

Number of selected features by estimated feature-relevance for predicting stroke in patients age 20-39 in females (left) and males (right).
List of Tables

3.1 Patient characteristics for validation cohort and lower risk subpopulations. All characteristics except age and BMI are reported as %. BMI = Body Mass Index; IQR = Interquartile Range; MI = Myocardial Infarction; TRS = TIMI Risk Score; EF = ejection fraction; BNP = B-type natriuretic peptide. -: data not available for the cohort.

3.2 Number of patients in validation cohort and low risk subgroups, and hazard ratio (HR) of risk metrics Morphologic Variability (MV) and MV in Beat-space (MVB). CVD = cardiovascular death; TRS = TIMI Risk Score; EF = left ventricular ejection fraction; BNP = B-type natriuretic peptide.

3.3 Univariable and multivariable hazard ratios (HR) of ECG-based risk metrics in the validation cohort. Metrics with significant multivariable HRs are in bold.

3.4 Univariable and multivariable hazard ratios (HR) of ECG-based risk metrics in the lower risk subgroup, TRS ≤ 4. Metrics with significant multivariable HRs are in bold.

3.5 Correlation of MVB and MV with other risk metrics (top section) and patient factors (bottom section). To normalize for the different numerical ranges of the different metrics, continuous risk metrics were dichotomized at the upper quartile in the placebo population. For categorical risk metrics with more than two categories, the highest risk categories were used: HRT = 2 and TRS ≥ 5.

3.6 Performance of MVB using segments of the ECG.
4.1 Datasets and Patient Characteristics. CVD = Cardiovascular Death, IQR = Interquartile Range (the values at the 25th and 75th percentiles, MI = Myocardial Infarction.

4.2 Area Under Curve (AUC) of time- and beat-frequency HRV measures averaged over 1,000 test sets. Bold indicates the higher AUC in each row, and * indicates the highest AUC.

4.3 AUC of WHRV models evaluated on holdout sets D2 and D3.

4.4 Hazard Ratio (HR) averaged over 1,000 test sets. 95% confidence intervals (CI) reported in parenthesis. Bold indicates the highest HR in each row (p<0.001 for unadjusted and adjusted for TRS; in the last row only WHRV in beat-frequency has a CI that does not include 1). *
*: In 7 out of 1,000 test sets, less than 30% of the patients had measured values of both EF and BNP, and therefore these test sets were excluded.

4.5 Final machine learning model in beat-frequency. Only features that are selected by the machine learning algorithm are shown. To compute the WHRV risk metric for a new patient, each feature is converted to a z-score by subtracting the mean and dividing by the standard deviation. The final risk metric is the sum of the products of the z-scores and the corresponding weights.

4.6 Most highly correlated time-frequency bands to the two learned beat-frequency bands. The time-frequency bands show a trend towards higher frequencies at higher heart rates, and the explained variances are almost always less than 67%.

5.1 Examples of Billing Codes.

5.2 Definitions of outcomes and keyw ord used in estimating the relevance of each feature. CVA= cerebrovascular accident (stroke); CHF= congestive heart failure; AMI= acute myocardial infarction; DM= diabetes mellitus; HCh= hypercholesterolemia.
5.3 Prediction tasks and number of patients. Acronyms as per Table 5.2
Chapter 1

Introduction

In medicine, risk models are used to predict a patient’s risk of developing adverse outcomes. For example, a risk model may use characteristics such as a patient’s age, gender, and past medical history to predict the probability that they will develop a stroke in the next year. These risk predictions are useful for determining the most appropriate medical therapies for each patient. For example, anticoagulants are used for stroke prevention in patients with the heart condition, atrial fibrillation. However, anticoagulants cannot be prescribed to all patients because they may cause excessive bleeding [35]. Thus physicians use risk models such as [56] to predict a patient’s risk of developing a stroke as part of the decision to prescribe these medications.

Developing a risk model starts with collecting data and extracting patient features and outcomes from the data. Examples of simple features include numerical values that quantify patient characteristics, such as age, gender, and presence (1) or absence (0) of medical conditions. Examples of more complex features include mathematical functions of other features and expert-annotated characteristics of a medical image. Next a risk model can be learned using the features and outcomes. Medical knowledge, or more generally, domain specific knowledge, is often helpful in risk modeling. In this thesis, we develop more accurate risk models by using novel approaches of leveraging domain knowledge to improve feature engineering and estimating feature relevance.

Feature engineering is the process of extracting features that lead to high model accuracy. For example, the body mass index (BMI) is defined as the patient’s weight
(in kg) divided by the squared height (in m). Because this index normalizes the weight by the estimated size of a patient, it may provide a better summary of the patient’s weight and health status relative to using the weight alone. Thus the BMI is a non-linear function of two other features, height and weight. Other functions can utilize more variables and be significantly more complicated, such as equations for estimating kidney function from blood tests [92]. In other domains such as image classification, learning such functions automatically from millions of examples may be possible using approaches such as deep learning [51]. However, medical datasets are typically much smaller, rendering careful feature design more critical.

Estimating feature relevance can improve model accuracy by focusing the model learning process on the most relevant features. In one application, features that are believed to be irrelevant can be removed from the model, a process termed feature selection [43]. Feature selection reduces the number of features used in the machine learning model and can aid interpretation of the model by experts, improve prediction performance by reducing the possibility of overfitting to irrelevant features, and reduce the computational time required to train models. For example, when trying to predict the future incidence of a heart attack, knowledge about human physiology may prompt one to use features related to heart, kidney, and lung, and ignore features related to past history of traffic accidents. In addition to selecting the best features as input to the risk model, some learning algorithms can take as input both relevant and irrelevant features, and select the most predictive features as part of modeling process.

In this thesis, we explore the utility of a novel frequency domain in engineering features from the heart’s electrical signal, the electrocardiogram (ECG). This new frequency domain is designed to take into account the inherent beat-to-beat variability of the heart rate. Next, we develop an automated approach to transfer medical knowledge extracted from text to improve the development of predictive models using health insurance records.
1.1 Applying Medical Knowledge to Aid Feature Engineering

In Chapters 3 and 4 we present two applications of applying medical knowledge to engineer features using the ECG. We use these features to predict whether a patient will die after an acute coronary syndrome (ACS), a class of disorders that includes heart attack, and show that our feature engineering approach improves prediction performance.

1.1.1 Motivation

ACS is an important problem because of its prevalence, relation to mortality, and associated economic burden. Each year, 1.1 million ACS-related hospitalizations occur in the United States alone, causing 110,000 deaths [68]. The total annual direct medical costs of these patients are estimated at $75 billion [100]. Although there has been significant progress in treatment of ACS patients, there is still room for improvement.

Predictive models are of interest in ACS patients because higher-risk patients derive a greater benefit from the antiplatelet drug tirofiban [66] and early invasive medical procedures such as catheterization and revascularization [26]. In addition, these studies also found that in certain groups of low-risk patients, these therapies were not associated with better outcomes. Together, these results suggest that these treatments should be targeted to high-risk patients, potentially leading to better outcomes and cost savings.

1.1.2 Contributions

We approach the problem of building risk models using feature engineering techniques to improve prediction performance. There are many ways to extract features from the ECG. A commonly used approach is frequency domain analysis, which as the name suggests, quantifies repeating activity of the ECG. This makes sense because
the heart cycle is quasi-periodic. Traditionally, frequency is measured in units of Hz, or per second. However, patients have different average heart rates, and a given patient’s heart rate varies over time and is heavily influenced by factors such as physical activity and stress. Consider a scenario with two patients, A and B, with average heart rates of 60 and 120 beats per minute (bpm), respectively. 0.5 Hz in the traditional frequency domain is equivalent to once every two heartbeats in patient A but once every four heartbeats in patient B. Thus this traditional “time-frequency” domain may measure different phenomenon at different heart rates.

Correspondingly, if an event is expected to repeat every two heartbeats, this would be equivalent to 0.5 Hz in patient A, but 1 Hz in patient B. The presence of such a repeating pattern would thus be more consistently captured in an alternative frequency domain, “beat-frequency,” which quantifies repeating activity with respect to heartbeats. We examine the hypothesis that risk models can be improved by extracting features in beat-frequency instead of time-frequency. Specifically, we show that:

- Using features in beat-frequency to build predictive models improves accuracy relative to time-frequency. This applies to both features based on ECG morphology (Chapter 3) and features based on the heart rate (Chapter 4).

- When we train models on randomly selected groups of patients, the same beat-frequency features are consistently found to be predictive. However, the features identified to be predictive in time-frequency are less consistent and vary depending on the set of patients the model was trained on. The higher consistency of beat-frequency is important because it allows easier interpretation of the predictive features, and enables the model to be reliably applied to future patients.

- When the heart rate increases (e.g., from 75 to 100 bpm), the time-frequency that certain events occur at also increases (e.g., from 0.25 to 0.29 Hz). This “shift” does not occur in beat-frequency. These observations may explain why different patients have different predictive features in time-frequency, but similar predictive features in beat-frequency.
1.2 Transferring Medical Knowledge From Text

In Chapter 5, we develop an automated approach to transfer knowledge from text articles written by biomedical experts to improve the accuracy of risk models when training data is scarce.

1.2.1 Motivation

Training models with small datasets is a particularly important problem in medicine because adverse outcomes may (fortunately) be rare. One of our prediction tasks in Chapter 5 is to predict the new onset of stroke. In our dataset, among 170,000 females aged 20-39, only 100 (0.1%) developed stroke within five years. When using logistic regression to model risk, a rule of thumb is to have 10 of the minority class (stroke in this case) per feature used [75]. Because we have \( \approx 10,000 \) features, it is easy to see that the amount of training data is inadequate to learn accurate models.

This problem is compounded if the original dataset is small. For example, a hospital may need a risk model to predict patients’ risk of acquiring an infection. The risk model needs to be tailored to that specific hospital because important features may be hospital specific. For example, infections can spread between patients by being in physical proximity, or having stayed in the same room. In [102], the authors used data from 3 hospitals ranging from 10,000 to 40,000 admissions. Despite a higher rate of adverse events (1%) relative to our stroke example, there were only 100 to 400 adverse outcomes.

In contrast to Chapters 3 and 4 where we focus on predicting adverse outcomes after an ACS, in Chapter 5, our goal is to develop a more general method that can be applied to many diseases and outcomes. As such, we evaluate our method on its ability to predict the new onset of a variety of diseases in several patient populations. We focus on five different cardiovascular diseases: cerebrovascular accident (stroke), congestive heart failure, acute myocardial infarction (heart attack), diabetes mellitus (the more common form of diabetes), and hypercholesterolemia (high blood cholesterol). In addition for each disease, we examine two different age groups (20-39 and
40-59), and both genders. This results in 20 different prediction tasks.

1.2.2 Contributions

Our approach leverages two unexploited aspects of risk modeling: textual descriptions of features and the outcome of interest, and publicly available external text data. For example, feature A may indicate a past history of hypertension, feature B may indicate a history of asthma, and our outcome of interest may be stroke. Given these information, an expert may be able to tell us that feature A is expected to be far more relevant in predicting stroke than feature B. However, manual expert annotation of the expected relevance of thousands of features is infeasible.

We tackle this problem by using the knowledge contained in the medical literature (such as PubMed and PubMedCentral) to estimate the relevance of each feature to our outcome of interest. These databases of published biomedical articles contain studies of the pathophysiology of various medical conditions and the effect of various therapies and medications. We develop an approach to use this knowledge to assess the textual feature descriptions to estimate the relative relevance of each feature. We then use the relevance estimates to learn accurate risk models despite the lack of training data. Specifically, we use the relevance estimates to scale the regularization of each feature. Regularization is a process that prevents models from learning weights that are too specific to the training data, thus improving generalizability of the risk model to new data. Our contributions are:

- We develop a method to estimate the relevance of each feature to the outcome of interest using models trained on large biomedical text databases.

- We use these relevance estimates to rescale features. Equivalently, more relevant features face weaker regularization.

- We show that our method improves the accuracy of risk models, and allows accurate predictions even when trained on datasets containing only 50 positive examples.
1.3 Organization of Thesis

The rest of the thesis is arranged as follows: Chapter 2 covers the background on the data that we use and predictive models that have been developed using similar data. The next two chapters explore the application of beat-frequency to features extracted to quantify ECG morphological changes (Chapter 3) and heart rate (Chapter 4). Chapter 5 demonstrates a method to incorporate knowledge from biomedical articles to improving prediction models developed on small datasets. Finally in Chapter 6 we summarize our findings and their implications and propose follow-up work.
Chapter 2

Background

In this chapter, we review the background for Chapters 3 and 4 in Section 2.1 and the background for Chapter 5 in Section 2.2.

2.1 Risk Stratification using Electrocardiograms (ECG)

The section focuses on background relevant for understanding our work on engineering features using the ECG. We start with a discussion of the heart and its electrical signal, the ECG. Next, we describe our disease of interest, acute coronary syndrome, and current methods for risk stratification.

2.1.1 The Heart and ECG

Here, we cover the basic facts about the heart; interested readers are referred to [55] for more information. The heart is a muscular organ that pumps blood around the body. The pumping action is driven by contractions of the myocardium, or heart muscle. This process of blood circulation removes waste products from and supplies nutrients and oxygen to various parts of the body. The heart is divided into two upper chambers (atria) and two lower chambers (ventricles) (Figure 2-1). The atria function to pump blood into the ventricles, while the ventricles pump blood out of the heart. Four one-way valves, one at the outlet of each of the four chambers ensure that
blood flows in a fixed direction. Blood travels in the following path: right atrium, right ventricle, lungs, left atrium, left ventricle, body, and back to the right atrium. “Body” refers to any organ, such as the brain, kidneys, and the heart itself (via the coronary arteries).

Cardiac contractions are driven by electrical impulses that travel across the heart using pathways illustrated in Figure 2-1. These impulses originate in an area in the right atrium, the sinoatrial node (SAN). The SAN contains pacemaker cells that initiate each heartbeat by firing an electrical impulse. The impulse spreads over the heart, driving coordinated contraction. First, the impulse spreads across the atrium, triggering atrial contraction. Next, the impulse reaches the atroventricular node (AVN) and is delayed momentarily. This delay allows for complete contraction of the atria, which ensures that the ventricles are filled with blood before they contract. In the next step, the impulse conducts along the specialized pathways: the His bundle, left and right bundle branches, and the Purkinje fibers to initiate ventricular contraction. At the same time, the atria relax. Finally, the ventricles relax. The electrical events corresponding to contraction and relaxation are called depolarization (triggered by impulse arrival) and repolarization (occurs automatically after depolarization), respectively.
Figure 2-2: A normal ECG with labels on characteristic features. Figure adapted from [1].

The coordinated electrical activities can be measured by the electrocardiogram (ECG). The ECG uses surface electrodes to measure the potential difference between standardized positions of the body’s surface. The measured signal is comprised of a characteristic repeating pattern (Figure 2-2), labeled with the letters P, Q, R, S, and T. The small P wave indicates atrial depolarization, the high amplitude QRS complex indicates ventricular depolarization, and the T wave indicates ventricular repolarization. The ventricular events have larger amplitudes because the ventricles are larger and have greater muscle mass. Atrial repolarization is typically hidden in the larger amplitude of the QRS complex.

At rest, an adult human heart typically beats at 60 to 100 beats per minute. Under conditions such as exercise, this rate can be substantially higher and is termed tachycardia. However, the heart rate is typically not constant. For example, even for a person at rest, the instantaneous heart rate will vary from slightly from beat to beat, as illustrated in Figure 2-3.

The heart rate is primarily modulated by the sympathetic and parasympathetic nervous systems [70]. Stimulation of the sympathetic branch accelerates heart rate, and stimulation of the parasympathetic branch slows down heart rate. Withdrawal of stimulation of the respective nervous systems have opposite effects. We will review the study of Heart Rate Variability (HRV) in Section 2.1.4. This variability in heart rate has implications in extracting features from the ECG, as we will see in Chapters 3.
2.1.2 Acute Coronary Syndrome (ACS)

In this section, we review background on ACS. For more details, [8] reviews pathogenesis of the disorder and treatment guidelines for healthcare providers.

As the term “acute” suggests, an ACS is a sudden cardiac event. Specifically, an acute coronary syndrome occurs when there is a mismatch between oxygen supply and demand in the heart muscles, or myocardium. This results in insufficient oxygen in the myocardium, or ischemia. The patient may experience symptoms such as pain in the chest, arm or jaw; chest pressure (“an elephant sitting on my chest”); diaphoresis (sweating); shortness of breath; and a “feeling of doom.” All or none of these symptoms may be present, and the latter case is termed a silent attack. A silent attack may be diagnosed post hoc based on findings such as scarring in the heart or ECG Q-wave changes.

When the patient symptoms and events leading up to the event are consistent with an ACS, physicians use a decision tree similar to Figure 2-4 to classify patients as having one type of ACS or another. In the first branch point, the physician may
ask if there were changes in the pattern of symptoms, such as new chest pain without physical activity, or more severe chest pain with the same level of physical activity. Next, the physician may use short (≈10 seconds) recordings of the ECG to determine if there are any changes in the ECG such as ST segment elevation. The ST segment is the part of the ECG signal between the S and T waves (Figure 2-2) in a single heart beat, and changes (either elevation or depression) indicate active heart tissue ischemia. If the patient has changes in symptoms, and no ST elevation, the patient is defined to have non-ST-elevation ACS (NSTEACS). The next branch point relies on drawing blood to check for elevation of certain biomarkers in the blood. Because the blood tests require more time compared to recording short ECG segments, blood is drawn and sent to the laboratory for testing at this point. However, treatment may be initiated if necessary before the blood test results are available.

In Chapters 3 and 4 we build risk models to predict the risk of death within a pre-defined time period (90 days or 1 year) in patients after a NSTEACS. We focus on NSTEACS because patients with the ST-elevation form are considered higher risk, and require invasive therapies within 90 minutes if possible [71]. Thus there is a greater need for risk prediction in the NSTEACS population. Furthermore, the proportion of patients with non-ST-elevation have increased, from 53% in 1999 to 77% in 2008 [104].

2.1.3 Clinical Risk Metrics

After patients are determined to have a NSTEACS, various methods are used to determine their risk of future adverse events, such as ischemia, another ACS, and death. For example, risk scores such as Global Registry of Acute Coronary Events (GRACE) [42] and Thrombolysis In Myocardial Infarction (TIMI) [9] produce risk estimates using a small number of clinical variables. The TIMI risk score starts at 0 and is increased by one for the presence of each additional risk factor (out of 7):

- Age ≥ 65
- ≥ 3 risk coronary artery disease (CAD) risk factors (family history of CAD, hypertension, hypercholesterolemia, diabetes, or current smoker).
- Significant coronary artery stenosis (narrowing, e.g., \(\geq 50\%\))
- ST deviation \(\geq 0.5\) mV
- Severe angina (chest pain, \(\geq 2\) episodes in past 24 hours)
- Use of aspirin in past 7 days
- Elevated blood biomarkers (creatine kinase MB fraction and/or cardiac specific troponin)

A physician can consult a risk table to determine the patient’s expected level of risk. For example, patients with TIMI risk score \(\geq 5\) have \(>26\%\) chance of death, myocardial infarction or severe ischemia requiring invasive therapy in the next 14 days \[9\]. The GRACE risk score incorporates more variables, but the computation is more complicated and requires specialized calculators.

Other measurements such as left ventricular ejection fraction (LVEF, or EF) and B-type natriuretic peptide (BNP) \[32\] are also used to risk stratify patients. The EF is defined as the fraction of blood in the left ventricle that is pumped out with each heart beat. A low EF indicates a decreased ability of the heart to pump blood. The EF is measured using an ultrasound device, the echocardiogram, and requires a specialized technician to conduct the test and a trained physician to interpret the output. An EF \(< 40\%\) is frequently used as a threshold for high risk. BNP is a biomarker that is released into the bloodstream in response to excessive ventricular stretch, and has been found to be elevated after myocardial ischemia, or a lack of oxygen \[32\]. BNP \(> 80pg/ml\) has been used as a threshold for high risk.

The ECG is also used in clinical risk stratification. The most common type of ECG recording is known as a 12-lead. The 12-lead captures the electrical activity of the heart from 12 different perspectives, and is about 10 seconds long. Each perspective is termed a “lead,” and multiple leads are helpful in ensuring that abnormalities observed are real and not noise. Figure 2-5 shows a typical 12-lead ECG. From left to right, four leads are plotted simultaneously. In the top three plots, these leads change approximately every three seconds. For example, the top signal shows the signals for leads I, AVR, V1, and V4 respectively. The bottom signal shows a continuous recording of lead II for the whole duration.
These recordings are used to diagnose NSTEACS, monitor ischemia, and to narrow down the source of observed abnormalities. Observations such as ST segment depression and T wave inversion are used in risk assessment \[8\], and ST segment depression is one of the components of the TIMI Risk Score.

In this thesis, we will be comparing our proposed methods relative to these clinical measures: TIMI Risk Score, EF, and BNP.

2.1.4 Long-term ECG Risk Metrics

In contrast to the relatively short 12-lead ECG, our work focuses on leveraging day-long Holter recordings. Risk metrics derived from these long ECG recordings have been shown to be associated with adverse outcomes in many studies. These ECG-based metrics can be broadly divided into ones that analyze heart rate changes, and ones that analyze changes in morphology.

**Heart Rate Variability (HRV)**

Heart rate based metrics are meant to quantify modulation of heart rate by the sympathetic and parasympathetic branches of the nervous system. The most established among them is a collection of metrics collectively termed HRV \([70]\). Conventional HRV metrics quantify the variability of the intervals between adjacent heartbeats in
milliseconds (ms). These intervals are termed normal-normal (NN) intervals because normal heartbeats are analyzed. In this thesis, we will compare our proposed measure with several of the most established ones:

- Standard Deviation of NN intervals (SDNN).
- Average Standard Deviation of NN intervals (ASDNN): the average of the standard deviation of NN intervals in all five minute segments in a day.
- Standard Deviation of Average NN intervals (SDANN): the standard deviation of the average of NN intervals in all five minute segments in a day.
- Heart Rate Variability triangular Index (HRVI): after computing a histogram of NN intervals, the HRVI is the maximum count in any bin of the histogram divided by the total number of NN intervals. In our work the bin size is 1/128s based on the sampling rate of our ECG (128Hz).
- Proportion of consecutive NN intervals that differ by more than 50 ms (PNN50).
- Root Mean Square of Successive Differences (RMSSD): after computing the time series of the difference between consecutive NN intervals, take the square of all the values, average them, and take the square root.
- Low Frequency / High Frequency (LF/HF): frequency domain measure that quantifies the ratio of energy in the low frequency band (0.04-0.15Hz) to that in the high frequency band (0.15-0.40Hz). In our work we compute the LF/HF value for each five-minute segment in a day, and take the median to be the final LF/HF value. We build on this metric in Chapter 4.

**Heart Rate Turbulence (HRT)**

HRT [84] quantifies the rate at which the heart rate returns to normal after a premature ventricular contraction (PVC). During a PVC, the ventricles contracts abnormally early before they have had time to fill fully, resulting in a weaker pulse. This activates compensatory mechanisms that initially accelerate the heart rate to preserve
blood pressure, and later decelerates it back to baseline. HRT has two components, turbulence onset (TO) and turbulence slope (TS). TO is defined as the difference between the sum of the two RR intervals after the PVC and two RR intervals before, divided by the sum of the two before:

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

TS is defined as the maximum positive slope of linear regression lines fitted to any sequence of five consecutive RR intervals in the 20 RR intervals following the PVC. Thus TO measures the initial acceleration of heart rate and TS measures the late deceleration. TO $\geq 0\%$ and TS $\leq 2.5\text{ms/RR interval}$ are two criteria associated with high risk. HRT is then defined as 2 (highest risk) if a patient meets both criteria, 1 (moderate risk) if a patient meets a single criteria, and 0 (lowest risk) otherwise. In two studies of myocardial infarction patients (577 and 614 patients respectively), $HRT = 2$ was a powerful risk predictor (risk ratio 3.2) of death comparable to EF $<30\%$ (risk ratio 2.9) [84].

**Deceleration Capacity (DC)**

DC [15] measures the average deceleration of heart rate. First, we find anchors, which are defined as RR intervals longer than the preceding interval. Next, segments of 4 RR intervals around the anchors are aligned and averaged:

$$X(n) = \frac{1}{N} \sum_{i=1}^{N} RR_{anchor_i+n}$$

where N is the number of anchors and $anchor_i$ indexes over the anchors. Thus $X(0)$ is the average of all the anchors and $X(-1)$ is the average of all the RR intervals before the anchors. DC is then defined as the average increase in RR interval between the two beats before the anchor and the anchor and the beat after the anchor:

$$DC = \frac{X(0) + X(1) - X(-1) - X(-2)}{4}$$
In two datasets of myocardial infarction patients (656 and 600 patients respectively), DC≤2.5ms was found to be a powerful predictor of mortality in patients with preserved EF (EF>30%) \[13\].

Severe Autonomic Failure (SAF)

SAF \[14\] is a combination of HRT and DC, and is defined as TO≥0%, TS≤2.5ms/RR interval, and DC≤2.5ms. In a study of 2343 myocardial infarction patients, patients that have preserved EF (>30%) but SAF have equivalent mortality relative to the patients with compromised EF (5-year mortality of 38% in both groups). Using both EF and SAF doubles the sensitivity relative to using EF alone (21% vs. 42%) of the risk model while preserving the mortality rate of predicted high risk patients.

T Wave Alternans (TWA)

Several ECG risk metrics quantify ECG morphology. Arguably the most established of these is TWA \[80\]. TWA evaluates the presence of an alternating pattern in the T wave: every other heartbeat has a higher (or lower) T wave. This has been hypothesized to be a marker of electrical instability in the heart.

There are two ways to measure TWA \[101\]. The Spectral Method and utilizes specialized equipment and sophisticated noise removal techniques to measure small (\(\mu V\)) beat-to-beat differences in T wave amplitude. This technique cannot be applied to more commonplace Holter data used in this thesis and will not be discussed further.

The second method, Modified Moving Average (MMA), can use data such as the Holter ECG. MMA uses a moving average of alternate beats. For example in a sequence of 6 beats: ABABAB, MMA averages the A and B beats separately to obtain two averaged beats \(\bar{A}\) and \(\bar{B}\). The biggest difference in the ST segment and T wave between \(\bar{A}\) and \(\bar{B}\) is defined as the TWA. The TWA computed in this thesis utilizes this MMA approach. In a case control study of myocardial infarction patients (46 cases, 92 controls \[91\]), \(TWA > 47\mu V\) had a relative risk of 5.5 for cardiovascular deaths. As in several other studies, the analysis required expert verification of TWA computation.
Morphologic Variability (MV)

MV [93] is a recent risk metric that is designed to quantify the beat-to-beat variability in the whole ECG waveform, as opposed to the ST segment and T wave in TWA. The computation of MV involves first producing a Morphologic Distance (MD) time series, which is a single value for every heart beat in the original ECG signal. Each MD value quantifies the difference between the heartbeat $i$ and $i+1$. Next, we divide this MD time series into five minute windows, and for each window we compute the energy in a “diagnostic frequency” ($0.30 - 0.55 \text{Hz}$). The final MV value is the 90th percentile of the energies across all windows in the day. Details of this method are available at [93]. In a study of 4500 NSTEACS patients, $MV > 52.5$ had a hazard ratio of 2 even after adjusting for TIMI Risk Score, EF, and two other ECG metrics. We build on this metric in Chapter 3.

2.1.5 Evaluation and Comparison of Risk Metrics

As can be seen above, the evaluation measures used in studies of risk metrics can vary substantially, making it difficult to compare the performance of different risk metrics. Here, we briefly introduce several of the most common risk metrics. The relative risk is the ratio of the proportions of adverse events ($p$), in a high-risk group relative to a low-risk group. The odds ratio is defined as the ratio of the odds ($\frac{p}{1-p}$) of adverse events between the two groups. The hazard ratio quantifies the ratio of the rate of adverse events over time between the two groups, and thus requires the timing (e.g., number of days since the time of prediction) of each adverse event. These three ratios vary between 0 and infinity. A value of 1 indicates no difference between the two groups, and thus no predictive value of the risk metric used to define risk. Higher values indicate a risk metric with better discriminatory ability.

Another evaluation measure is the area under the receiver operating characteristic curve (AUC, or $c$-statistic). Given a patient who will experience the adverse event and one who will not, the AUC measures the probability of the risk metric correctly identifying the higher risk patient. The AUC ranges from 0 to 1, where 0.5 indicates
random performance and higher values indicate a risk metric with higher discriminatory performance. Unlike the ratios mentioned above, the AUC does not require the definition of a high-risk cutoff or threshold (e.g., \( \geq 5 \) is the threshold for the TIMI Risk Score). This property of the AUC makes it convenient during risk model development.

A few factors are critical in interpreting a risk metric. First, for evaluation criteria based on a single high-risk cutoff, the corresponding performance may change substantially based on the cutoff used. In clinical use, consideration in choosing the cutoff include the cost and side risk of interventions, and the absolute (in addition to the relative) risk of adverse events in high and low-risk groups.

Second, several evaluation measures can be adjusted for known risk factors such as age and hypertension. The adjustment allows inferences about the discriminatory performance of the risk metric that is independent of the adjusted factors. This is important because of correlations between risk metrics. For example, if a risk metric is highly correlated with age and has poor performance after adjusting for age, then no additional information has been gained from measuring or computing the risk metric. In general, adjusting for additional risk factors decreases the evaluation metric, and thus adjusted evaluation measures cannot be directly compared to unadjusted ones.

Finally, the patient population of each study needs to be taken into account. For example, the fact that a risk metric performs well in an elderly population may not indicate that it has similar predictive performance in a population of young athletes. Therefore, it is important for studies to report the population characteristics of their study cohort. Some studies further report performance in various subgroups. These data are important because some patients are already considered high risk. For example, patients with high EF may be considered relatively low risk, and may not be treated based on that risk factor. Therefore an accurate risk metric in this population would identify high-risk patients who may otherwise not receive needed therapy.
2.2 Risk Stratification using Health Insurance Records

The section focuses on background relevant for understanding our work on leveraging knowledge from text to improve accuracy of models built using small datasets Chapter 5. We will cover background on related machine learning methods, methods of modeling natural language, and the type of data that we work with.

2.2.1 Transfer Learning

Traditional machine learning uses a specific dataset to learn a specific model for the prediction task of interest. By contrast, transfer learning improves a machine learning model of interest, termed the “target,” by transferring knowledge from one or more external “sources.” For the purposes of this work, we will briefly review supervised transfer learning, where the target task is to predict a specific outcome. Interested readers are directly to [74] for a more thorough review. Throughout this section, we will use variants of the same example: our target task is to generate a predictive model for whether patients seen at Hospital T will die in the next year. Our source data are patients at Hospital S.

Transfer learning can be classified into several different types, depending on two factors of the source and target: the domain and task. “Domain” refers to the data: feature space and feature distribution. For example, if two hospitals record the same data in the same format, their feature spaces are identical. Feature spaces may
differ if we use hospital specific features such as the room that the patient was assigned to. Feature distributions are identical if the two datasets are expected to be similar, such as if the two hospitals see similar patient populations. Feature distributions may differ, for example, if one hospital specializes in diabetic patients, and the other specializes in cancer patients.

“Task” refers to the prediction task and mathematical function in the prediction model. The prediction tasks are identical, for example, if the goal is to predict the same outcome over the same time period in both hospitals. The mathematical function in the model is learned from the data. One way this can differ is if the two hospitals have different numbers of deaths.

Transfer learning is traditionally possible only if there are similarities between the domain or task. Examples of applications of transfer learning to medical applications include adapting global surgical models to individual hospitals [52], enhancing hospital-specific predictions of infections [102], and improving surgical models using data from other surgeries [41]. In this work, we propose a method to transfer knowledge when the domains and task are both different.

Intuitively, our source domain/task is to learn a representation of the semantic meaning of words used in the medical literature. We will review the model we use and other similar models in the following section, 2.2.2. Our target domain/task is to predict the new onset of a specific disease using a patient’s past medical history. Our data comes from an health insurance database, where each feature represents a standardized billing code (diagnosis, procedures, or medications). We will review type of data in Section 2.2.5.

Thus our source and task were unrelated in both type of data (text versus structured medical data) and task (learning word meaning versus predicting the presence or absence of a disease in the future). To our knowledge we demonstrate the first instance of transfer knowledge between such unrelated tasks.
2.2.2 Distributed Representation of Words

Models that characterize word sequences in natural language are termed language models. Traditionally, people used a statistical approach based on n-gram [61]. These models predict the next word given the preceding $n-1$ words. In a 3-gram (trigram) model with $V$ distinct words for example, the model remembers the counts of all possible $V^3$ 3-word sequences, where $V$ may be $> 10^5$. These counts are learned from a training dataset, termed a corpus. In the simplest case, the probability of seeing “is” after “my name” is the number of times “my name is” occurs divided by the number of times “my name X” occurs, where $X$ is any word. Because this model captures word dependencies that extend to the previous $n-1$ words, the complexity of natural language that the model can represent grows with $n$. However, this method runs into the curse of dimensionality: the number of distinct n-gram word sequences grows exponentially. This is problematic because the model may encounter unseen $n-1$-word combinations and thus have no information with which to predict the $n^{th}$ word.

Another approach is the distributed, or continuous vector representation of words [17]. In this class of methods, each word is represented by a $d$-dimensional numerical vector that is learned from the data. For example, if a human was given the task of representing words with such a vector, each element may correspond to a property of the word, such as tense, gender (if the word is gender specific), and word type (such as noun, verb, etc). The model learns to represent words that appear in similar contexts with similar vectors. For example, “toad” and “frog” may have more similar vectors than “toad” and “chair.” Thus, the model may be able to make reasonable predictions about words being used in different contexts based on words that it has seen used in similar contexts. There have been many approaches to learning such representations, such as [29, 64, 99].

Because $d$ is typically only on the order of $10^3$, this distributed representation is much more compact than the naïve one-hot encoding. The one-hot encoding is a sparse vector that is the length of the vocabulary ($> 10^5$). Each position in the vector
represents a distinct word in the vocabulary, and only the position corresponding to the input word is one. This reduces the number of parameters that needs to be learned when words are used as input to a machine learning algorithm.

2.2.3 Word2vec

In this work, we leverage a relatively recent approach, word2vec [62, 63]. There are two architectures in this method: continuous bag of words (CBOW) and skip-gram. We will very briefly review these two architectures here. Readers are directed to [81, 39] for approachable mathematical derivations of the method.

Both CBOW and skip-gram are shallow neural networks (Figure 2-7). Both methods take as input one or more words in a one-hot encoding. Similarly, the prediction output is another word in the one-hot encoding format. CBOW uses the context, or the surrounding words to predict the word of interest. By contrast, skip-gram uses the word of interest to predict the words surrounding it. In both cases, the model learns an unique vector of real numbers for each word.

Word2vec models computed using the skip-gram approach are derived by implicitly factorizing a word-context matrix [53]. Each cell in this matrix is computed from the co-occurrence of the word and its context using pointwise mutual information. The authors of [53] showed that singular value decomposition (SVD) can be applied
to this matrix to extract vectors with similar properties to word2vec vectors.

2.2.4 Word Vector Properties

The learned word vectors have several interesting properties. Empirically, words with similar meanings have similar vectors. These similarities are measured by taking the dot product of the normalized word vectors, where the normalization sets the $l^2$ norm ($\sqrt{\sum x_i^2}$) of each vector to 1. The similarities fall in the range $[-1,1]$, where 1 indicates perfect similarity and 0 indicates low similarity. For example, “toad” and “frog” have higher similarity values than “toad” and “chair.”

Second, linear operations on the word vectors encode semantic relationships. For example, vector(‘king’) - vector(‘queen’) is approximately equal to vector(‘man’) - vector(‘woman’). Other vector operations apply: vector(‘king’) - vector(‘man’) + vector(‘woman’) is approximately vector(‘queen’). Another example is the country-capital relationship: vector(‘Athens’) - vector(‘Greece’) is approximately equal to vector(‘Beijing’) - vector(‘China’). A list of such word pairings that are used to test the models is available at [5].

In Chapter 5, we utilize the dot product similarity measure to quantify the similarities between feature descriptions and the outcome of interest.

2.2.5 Health Insurance Records

Health insurance data are used for billing purposes. When an insured patient gets treated or is seen by a healthcare provider, most of the costs are borne by an external party (the insurance provider). To standardize payments, billing codes such as International Classification of Diseases 9th Revision, Clinical Modification (ICD-9) [6] are used.

Although there are many types of billing codes, here we will review the ones that are present in our dataset: ICD-9 and the medication billing coding system.

\footnote{Similar to how shoppers frequently pay less than the “manufacturer recommended price” for an item, the actual amount paid to different institutions for the same billing code may vary based on negotiations. However, billing codes are necessary to standardize terminology.}
ICD-9 Codes

ICD-9 codes [6] were developed by the World Health Organization to standardize disease classification. This enables epidemiological studies of diseases despite the use of different languages or terminology by different physicians or institutions. For example, the code 250 is standardized to represent diabetes mellitus, the more common form of diabetes.

In our dataset, there are two types of ICD-9 codes, diagnoses and procedures. Diagnosis codes follow the XXX.XX format: three digit codes with two digits after a decimal point. There are also two sets of supplementary codes. The first set starts with “V”: VXXX.XX and are used for people who are not currently sick, such as for organ donation and vaccinations, and people who require long-term treatment, such as dialysis. The second set starts with “E”: EXXX.XX and identifies the cause of external injuries or poisoning, such as type of motor vehicle accident. ICD-9 procedure codes follow the XX.XX format and are used to indicate the type of operation or procedure that the patient received.

All of the types of ICD-9 codes mentioned above are arranged as hierarchies, as

Figure 2-8: Example of hierarchy for ICD-9 diagnosis codes (left) and ATC medication codes (right).

Anatomical Therapeutic Chemical (ATC) Classification System [7]. Billing data are also termed “administrative data.”
illustrated in Figure 2-8. For example, 410 represents acute myocardial infarction, or heart attack. In this case (but not for all codes), the fourth digit indicates the specific part of the heart that was damaged, and the fifth digit indicates if this was the first time that the patient was seen for this episode of heart attack.

Going one level up, the digits before the decimal point can be grouped into categories. For example, ICD-9 codes that fall between 390 and 459 are all types of circulatory system diseases. Codes that fall between 410 and 414 are ischemic heart diseases.

Frequently, not all of the digits are used. This is particularly true of digits after the decimal point. This can be due to insufficient information at the time of coding, such as unknown part of the heart that was implicated in a heart attack, or human error.

**ATC Codes**

ATC codes [7] were also developed by the World Health Organization, and classify drugs according to the organ system that they act on, the class of medication that they fall under, and their chemical properties.

Like the ICD-9, ATC codes are also hierarchical. The ATC hierarchy contains five levels (Figure 2-8). The first level contains a single letter and describes the organ system that the drug acts on. Examples include alimentary tract and metabolism (A), cardiovascular system (C), nervous system (N), and respiratory system (R). The second level contains two digits and indicates the general class of drug. Examples include diuretics (C03, removes water from the body and is used in conditions such as heart failure), and beta blockers (C07, slows down heart rate). The third level contains one letter and is a more specific class than the second. For example, diuretics can be divided into low ceiling (C03A) and high ceiling (C03B), which refer to whether the drug effects can be increased with higher dosage. The forth level consists of one letter and indicates a more specific type of drug. For example, high-ceiling diuretics can be divided into plain sulfonamides (C03CA), aryloxyacetic acid derivatives (C03CC), etc. Finally, the fifth level contains two digits and indicates the specific drug, such as
furosemide (C03CA01) and bumetanide (C03CA02).

Unlike the ICD-9 codes, the ATC codes are almost always complete, with no missing trailing letters or digits. This is because drug prescriptions have to be precise with respect to the drug name. In rare instances, the ATC codes are incomplete in our dataset. This occurs when the drug used has not been assigned a specific ATC code, and therefore the mapping stops at a higher level.
Chapter 3

Risk Stratification Using ECG:
Morphological Variability in Beat-space

3.1 Introduction

Risk stratification after acute coronary syndrome (ACS) involves integrating a diverse array of clinical information. Risk scores such as the Global Registry of Acute Coronary Events (GRACE) [42] and Thrombolysis In Myocardial Infarction (TIMI) [9] scores aid in this process by incorporating clinical information such as cardiac risk factors and biomarker data. Unfortunately, existing metrics like these only identify a subset of high-risk patients. For example, the top two deciles of the GRACE score and a high TIMI Risk Score (≥ 5) captured 67% and 40% of the deaths, respectively. That a significant number of deaths will occur in populations that are not traditionally considered to be high risk highlights a need for tools to discriminate risk further [40]. In this regard, the use of computational biomarkers may provide additional information that could improve our ability to identify high risk patient subgroups [94]. Indeed, several studies showed that ECG-derived computational metrics significantly improve the ability to risk stratify in subgroups with relatively preserved left
ventricular ejection fraction (EF) [12, 13, 14, 94].

An overview of these risk measures can be found in Section 2.1.4. Briefly, ECG-based metrics can be broadly divided into ones that analyze heart rate changes, and ones that analyze changes in morphology. Examples of heart rate-based metrics include Heart Rate Variability (HRV) [70], Deceleration Capacity (DC) [15], Heart Rate Turbulence (HRT) [84], and Severe Autonomic Failure (SAF) [14]. Morphology based metrics include T Wave Alternans (TWA) [80], which is designed to measure specific alternating changes in cardiac repolarization and Morphologic Variability (MV) [93], which is designed to quantify the beat-to-beat morphologic variability in ECG signals.

A key parameter in MV is the diagnostic frequency band of 0.30-0.55Hz. Higher variability in this frequency band is associated with adverse outcomes such as death [93]. In this study, we evaluate the hypothesis that measuring frequency in terms of cardiac cycles (“beat-frequency”) will result in a more accurate risk metric. We term this new metric MV in Beat-space (MVB), and compare MVB with clinical and other ECG risk metrics.

### 3.2 Methods

#### 3.2.1 Dataset

We used two datasets of ECG recordings in this work, a derivation [25] and a validation cohort [67], from two clinical trials of patients with non-ST-elevation ACS (NSTEACS). Patient characteristics of our cohorts are shown in Table 3.1. The derivation cohort consisted of 765 patients, with 14 cardiovascular deaths (CVD) (1.8%) within the median followup period of 90 days. We used the derivation cohort to derive parameters for a new morphologic variability risk metric, MVB, described below). This is the same population that the original Morphologic Variability metric was derived from [93].

We then tested the ability of each ECG-based metric to identify high risk patients on the validation cohort. In order to compare our proposed metric with established
clinical measures, we only included patients who had measured values for both the left ventricular ejection fraction (EF) and B-type natriuretic peptide (BNP). This population included 1082 patients, with 45 CVD within the median followup period of one year (4.5%).

In addition, we evaluated the performance of these metrics on several subgroups of patients that are considered “low-risk” based on the TIMI Risk Score (TRS), EF, and BNP. These low risk criteria are listed in the first column of Table 3.2. We did not use the GRACE risk score because we did not have all the variables required to compute it. The number of patients and CVD in these populations are presented in Table 3.2, sorted by decreasing one year CVD rates. Patient characteristics of representative low-risk subgroups (TRS ≤ 4 and the combination of TRS ≤ 4 and BNP ≤ 80) are shown in Table 3.1.
Table 3.2: Number of patients in validation cohort and low risk subgroups, and hazard ratio (HR) of risk metrics Morphologic Variability (MV) and MV in Beat-space (MVB). CVD = cardiovascular death; TRS = TIMI Risk Score; EF = left ventricular ejection fraction; BNP = B-type natriuretic peptide.

<table>
<thead>
<tr>
<th>Population</th>
<th># patients (CVD rate)</th>
<th>HR adjusted for:</th>
<th>MVB HR, p</th>
<th>MV HR, p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validation Cohort</td>
<td>1082 (4.5%)</td>
<td>TRS, EF, BNP</td>
<td>2.1, 0.016</td>
<td>2.6, 0.002</td>
</tr>
<tr>
<td>TRS ≤ 4</td>
<td>864 (2.7%)</td>
<td>EF, BNP</td>
<td>3.0, 0.014</td>
<td>3.7, 0.003</td>
</tr>
<tr>
<td>TRS ≤ 4, EF &gt; 40</td>
<td>776 (2.3%)</td>
<td>BNP</td>
<td>3.8, 0.007</td>
<td>3.0, 0.007</td>
</tr>
<tr>
<td>TRS ≤ 4, BNP ≤ 80</td>
<td>538 (1.6%)</td>
<td>EF</td>
<td>7.8, 0.014</td>
<td>4.4, 0.044</td>
</tr>
<tr>
<td>TRS ≤ 4, EF &gt; 40, BNP ≤ 80</td>
<td>503 (1.3%)</td>
<td>-</td>
<td>14.9, 0.014</td>
<td>5.9, 0.040</td>
</tr>
</tbody>
</table>

Figure 3-1: Because of instantaneous heart rate changes, cardiac events are periodic only with respect to heartbeats but not with respect to time.

3.2.2 Morphologic Variability in Beat-space (MVB)

Overview

MV measures beat-to-beat variability in ECG morphology in the diagnostic frequency of 0.30 to 0.55Hz. Taking the inverse of these two frequency values give the temporal periods 3.3s and 1.8s respectively, implying that high variability in beat-to-beat ECG morphology every two to three seconds is associated with death. However, because of beat-to-beat variation in heart rate, cardiac activity is periodic with respect to heart beats instead of with respect to time. For example, as illustrated in Figure 3-1, events observed at two-second intervals may correspond to different parts of the ECG waveform. However, by definition, events observed at two-beat intervals correspond...
Accordingly, we speculated that it might be useful to analyze frequency relative to heartbeats rather than relative to time. This changed the analysis space from time to beats, and consequently changed the frequency domain (units of Hz) to a “beat-frequency” domain (units of cycles/beats).

This concept was first reported as beatquency [57], where it was applied to the heart rate time series to classify sleep stages. The authors found through visual inspection that the beat-frequency spectra were highly consistent in the same subject across different nights, and across different subjects. For ease of interpretation, we report the beat-frequency bands as their inverse, i.e., “every x beats.”

**Computation of MVB**

We calculated MVB using the first 24 hours of ECG for each patient. Many steps are similar to the MV computation [95]. The important steps are briefly described here, and contrasted with MV in Figure 3-2. We first preprocessed the ECG using the Signal Quality Index [54], to help ensure that only normal beats are studied. Next, we removed baseline wander by subtracting the median filtered signal [31]. Then,
we normalized the ECG signal amplitude across different patients by dividing each ECG signal by the mean R wave amplitude for that patient. For example, if the mean R wave amplitude of the patient is 1.5 mV, the entire ECG signal is divided by 1.5. Thus this rescaling procedure did not affect isoelectric segments (with zero amplitude), and scaled the rest of the signal such that the mean R wave amplitude is 1. This step corrects for calibration errors and inter-patient differences, and enables more meaningful comparison of morphological differences.

Next, we converted the ECG signal into a beat-to-beat variability time series termed the Morphologic Distance (MD) time series \[95\]. The MD time series quantifies beat-to-beat morphologic changes in the ECG signal; each MD point was defined by the sum of the squared differences between the aligned beats (Figure 3-2 Step 1). In the original MV risk metric, the MD time series had time in seconds as the x-axis. Here, we used the heartbeat index \((1,2,3,...)\) as the x-axis. When heartbeats were removed in the preprocessing step, the gap was estimated by the timings and heartbeat intervals (RR intervals) adjacent to the gap: \(\text{gap} = \text{round}\left(\frac{\text{gap in seconds}}{\text{average}(\text{RR before gap}, \text{RR after gap})}\right)\).

Next, we segmented the MD time series into five-minute windows. Similar to the original MV, we next estimated the power spectrum of each five-minute window using the Lomb Scargle periodogram (LSP) \[60, 28\]:

\[
P(\omega) = \frac{1}{2\sigma^2} \left( \frac{\left[\sum_i (x_i - \mu) \cos \omega (t_i - \tau)\right]^2}{\sum_i \cos^2 \omega (t_i - \tau)} + \frac{\left[\sum_i (x_i - \mu) \sin \omega (t_i - \tau)\right]^2}{\sum_i \sin^2 \omega (t_i - \tau)} \right)
\]

where \(t_i\) is the \(i^{th}\) timepoint (heartbeat index in MVB), \(x_i\) is the signal value at that timepoint, \(\mu\) is the mean of the signal values, and \(\sigma\) is the standard deviation of the signal. For a given value of \(\omega\), \(\tau\) is defined as:

\[
\tan 2\omega \tau = \frac{\sum_i \sin 2\omega t_i}{\sum_i \cos 2\omega t_i}
\]

Thus by utilizing both the timepoint \((t_i)\) in addition to the time series values \((x_i)\), the LSP provides a natural way to handle unevenly sampled data without using re-
sampling methods and their accompanying assumptions [65]. In the original MV, the LSP was used to handle both the irregular timings of heartbeats and gaps introduced by heartbeat removal in the preprocessing stage. Here, although heartbeats are at regular intervals, we used the LSP to handle the gaps and to ensure that the methods were comparable.

For each five-minute window, we then computed the energy in a diagnostic beat-frequency (Figure 3-2, Step 2). This diagnostic beat-frequency was optimized in the derivation cohort of patients (described below). The 90th percentile of these energies from all of the five-minute windows was defined to be the MVB for that patient (Figure 3-2, Step 3).

**Optimizing MVB Parameters**

We optimized the diagnostic beat-frequency for MVB in the derivation cohort in a manner similar to that for MV [95]. For every possible start and end beat-frequency over the range of every 2 beats to every 100 beats, we computed the 90th percentile of the energy in that beat-frequency band over all five-minute windows in the day. We then computed the area under the receiver operating characteristic curve (AUC) [44] for that diagnostic beat-frequency across all patients in the derivation cohort. The AUC is described in Section 2.1.3 but briefly, the AUC measures prediction performance and ranges from 0 to 1. 0.5 indicates random performance, and higher values are better. Finally, we chose the beat-frequency range with the highest AUC to be the optimal diagnostic beat-frequency. The AUCs are presented as a heat-map, where each point in the heat-map corresponds to the AUC for MVB computed using a specific beat-frequency range as indicated by the axes (Figure 3-3). We obtained an optimal diagnostic beat-frequency of every 2 to 7 beats, with an AUC of 0.725. This diagnostic beat-frequency is equivalent to 0.50 and 0.14 cycles/beat, respectively, or 0.50 and 0.14 Hz at 60 beats per minute (bpm), and 0.83 and 0.23 Hz at 100 bpm.
Figure 3-3: Optimizing the diagnostic beat-frequency for maximum AUC in the derivation cohort. Our peak AUC is 0.73, at every 2 to 7 beats. The inset illustrates the Receiver Operating Characteristic (ROC) curve for this optimal diagnostic beat-frequency.

3.2.3 Analysis of MVB in Specific ECG Segments

To gain some insight into the physiology underlying the variability measured by MVB, we computed MVB based on partial segments of the ECG. This is inspired by the fact that different segments of the ECG indicate different physiological activity. First, we preprocessed the ECG signal as per the steps in MVB. Next, we computed the beat-to-beat MD time series as before. However, this time we divided each ECG heartbeat into four segments of equal length. We then used the alignment to extract the partial distance assigned to each quarter segment, resulting in four values for each original MD value: MD1, MD2, MD3, and MD4; MD=MD1+MD2+MD3+MD4. This resulted in four new “partial MD” time series that summarized the variability
in partial segments of the ECG. The remaining steps of conversion to the frequency
domain, summing the energy in the range every 2 to 7 heartbeats, and taking the
90th percentile were identical.

Although taking four equal segments may assign the PQRST waves to different
segments, we started with this approach because all the waves except R are lower
amplitude and more difficult to label reliably. Moreover, patients’ ECGs have differ-
ences such as widened QRS, abnormally shaped T waves, and concave or convex ST
segments that further confound detectors. Thus we reasoned that this simple proce-
dure would provide evidence to guide further work. Our results indicated that more
careful segmentation into specific waves is unlikely to further improve performance.

3.2.4 Correlation among Time and Beat-frequency Features

To further elucidate the differences between time- and beat-frequency, we computed
the correlation between bands in both types of frequency domains. First, we com-
puted the 90th percentile of the energy in each frequency band across all five-minute
windows as done in MV. Next, we calculated the pairwise correlation between different
frequencies. We repeated these two steps for the beat-frequency bands computed in
MVB. These procedures indicated that beat-frequency could be more easily grouped
into bands with high intra-band correlation and low inter-band correlation for further
study.

3.2.5 Comparison with Published Risk Metrics

We compared MVB with other published risk metrics: Heart Rate Variability (HRV)
[70], Heart Rate Turbulence (HRT) [84], Deceleration Capacity (DC) [15], Severe
Autonomic Failure (SAF) [14], T-Wave Alternans (TWA, computed using a fully au-
tomated version of the modified moving average [101]), TIMI Risk Score (TRS) [9],
left ventricular ejection fraction (EF), and B-type natriuretic peptide (BNP) [32].
We computed the following HRV metrics: Low Frequency / High Frequency (LFHF),
Standard Deviation of Average NN intervals (SDANN), Heart Rate Variability trian-
gular Index (HRVI), Average Standard Deviation of NN intervals (ASDNN), Standard Deviation of NN intervals (SDNN), Proportion of consecutive NN intervals that differ by more than 50 ms (PNN50), Root Mean Square of Successive Differences (RMSSD) [70]. All ECG metrics were computed using previously described methods [94], and all metrics were dichotomized for statistical analysis, as described in the next section.

3.2.6 Statistical Analysis

We evaluated the clinical and ECG metrics based on univariable and multivariable [1]-year hazard ratios (HR). HRs quantify the number of adverse events per unit time. We computed HRs using the Cox proportional hazards regression model [30], a commonly used model that assumes that the ratio of adverse events per unit time in the high and low risk groups remain constant. In each population, patients at the highest quartile of risk were defined to be high risk for HR computation. This enabled meaningful comparisons of hazard ratios for two reasons: (1) none of the established cutoffs were derived for the specific populations analyzed here, and (2) this ensured equal proportions of patients in each metric’s high risk category. EF and BNP were dichotomized at $\leq 40\%$ and $> 80\text{pg/ml}$ [32], respectively. After dichotomization, each metric was replaced by a binary value; for example EF was 1 if EF $\leq 40\%$, and 0 otherwise. For categorical metrics with more than two categories (TRS and HRT), the highest risk categories were used: TRS $\geq 5$, HRT $= 2$. Each multivariable hazard ratio was computed by including TRS, EF, BNP, and a single ECG metric. In the low risk subgroups defined using one or more of the variables TRS, EF, and BNP, the respective variables were removed from the multivariable hazard ratio computation.

\footnote{Although “multivariable” and “multivariate” are often used interchangeably, they have distinct meanings [45]. “Multivariate” refers to predicting multiple outcomes (e.g., outcomes over time for each patient, or different types of adverse events). On the other hand, “multivariable” refers to using multiple \textit{input variables} to predict the outcome. For example, a multivariable model may use three features to predict a single outcome.}
Table 3.3: Univariable and multivariable hazard ratios (HR) of ECG-based risk metrics in the validation cohort. Metrics with significant multivariable HRs are in bold.

<table>
<thead>
<tr>
<th>Risk Metric</th>
<th>Univariable 1-year HR (95% CI)</th>
<th>p</th>
<th>Multivariable 1-year HR (95% CI) (adjusted for EF,BNP,TRS)</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRS</td>
<td>4.42 (2.46,7.92)</td>
<td>&lt;0.001</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>BNP</td>
<td>3.20 (1.70,6.02)</td>
<td>&lt;0.001</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>EF</td>
<td>2.76 (1.43,5.34)</td>
<td>0.003</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MV</td>
<td>3.26 (1.82,5.85)</td>
<td>&lt;0.001</td>
<td><strong>2.62 (1.43,4.81)</strong></td>
<td>0.002</td>
</tr>
<tr>
<td>DC</td>
<td>3.01 (1.68,5.41)</td>
<td>&lt;0.001</td>
<td><strong>2.26 (1.22,4.17)</strong></td>
<td>0.009</td>
</tr>
<tr>
<td>HRV-LF/HF</td>
<td>3.25 (1.81,5.83)</td>
<td>&lt;0.001</td>
<td><strong>2.21 (1.20,4.09)</strong></td>
<td>0.011</td>
</tr>
<tr>
<td>MVB</td>
<td>2.70 (1.50,4.85)</td>
<td>0.001</td>
<td><strong>2.11 (1.15,3.89)</strong></td>
<td>0.016</td>
</tr>
<tr>
<td>HRV-SDANN</td>
<td>1.84 (1.01,3.36)</td>
<td>0.048</td>
<td>1.53 (0.83,2.81)</td>
<td>0.171</td>
</tr>
<tr>
<td>HRV-HRVI</td>
<td>1.37 (0.73,2.57)</td>
<td>0.334</td>
<td>1.12 (0.59,2.13)</td>
<td>0.728</td>
</tr>
<tr>
<td>HRV-ASDNN</td>
<td>1.36 (0.72,2.55)</td>
<td>0.344</td>
<td>1.09 (0.58,2.07)</td>
<td>0.790</td>
</tr>
<tr>
<td>HRV-SDNN</td>
<td>1.21 (0.64,2.31)</td>
<td>0.558</td>
<td>1.03 (0.54,1.98)</td>
<td>0.925</td>
</tr>
<tr>
<td>TWA</td>
<td>1.11 (0.57,2.15)</td>
<td>0.758</td>
<td>0.82 (0.39,1.71)</td>
<td>0.597</td>
</tr>
<tr>
<td>HRT2</td>
<td>1.37 (0.58,3.25)</td>
<td>0.474</td>
<td>0.77 (0.32,1.89)</td>
<td>0.574</td>
</tr>
<tr>
<td>HRV-PNN50</td>
<td>0.75 (0.36,1.55)</td>
<td>0.432</td>
<td>0.76 (0.37,1.58)</td>
<td>0.459</td>
</tr>
<tr>
<td>HRV-RMSSD</td>
<td>0.64 (0.30,1.38)</td>
<td>0.255</td>
<td>0.75 (0.35,1.61)</td>
<td>0.452</td>
</tr>
<tr>
<td>SAF</td>
<td>1.18 (0.42,3.31)</td>
<td>0.750</td>
<td>0.68 (0.24,1.97)</td>
<td>0.481</td>
</tr>
</tbody>
</table>

3.3 Results

3.3.1 Association between Risk Metrics and CVD in the validation cohort.

In the univariable analysis, TRS, HRV, BNP, DC, EF, MVB, and HRT were all associated with the risk of CVD (Table 3.3). After adjusting for TRS, BNP and EF; only MV, DC, HRV-LFHF, and MVB remained significantly associated with CVD (HRs from 2.1 to 2.6, p<0.05 for all (Table 3.3)).

3.3.2 Association between Risk Metrics and CVD in low-risk subgroups

22 out of 45 CVD (49%) occurred in the subgroup with TRS≤4, and this subgroup had a 1-year CVD rate of 2.7%. In this population, BNP, EF, MV, MVB, HRV-
Table 3.4: Univariable and multivariable hazard ratios (HR) of ECG-based risk metrics in the lower risk subgroup, TRS ≤ 4. Metrics with significant multivariable HRs are in bold.

<table>
<thead>
<tr>
<th>Risk Metric</th>
<th>Univariable 1-year HR (95% CI)</th>
<th>p</th>
<th>Multivariable 1-year HR (95% CI, adjusted for EF, BNP)</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>BNP</td>
<td>2.93 (1.23, 6.98)</td>
<td>0.015</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>EF</td>
<td>2.66 (0.98, 7.22)</td>
<td>0.054</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MV</td>
<td>4.45 (1.90, 10.42)</td>
<td>0.001</td>
<td>3.67 (1.53, 8.79)</td>
<td>0.003</td>
</tr>
<tr>
<td>MVB</td>
<td>3.66 (1.58, 8.48)</td>
<td>0.002</td>
<td>2.99 (1.25, 7.15)</td>
<td>0.014</td>
</tr>
<tr>
<td>HRV-LFHF</td>
<td>3.01 (1.31, 6.95)</td>
<td>0.010</td>
<td>2.32 (0.98, 5.53)</td>
<td>0.057</td>
</tr>
<tr>
<td>DC</td>
<td>2.55 (1.10, 5.91)</td>
<td>0.029</td>
<td>1.97 (0.82, 4.73)</td>
<td>0.132</td>
</tr>
<tr>
<td>HRT2</td>
<td>2.56 (0.86, 7.67)</td>
<td>0.092</td>
<td>1.85 (0.60, 5.74)</td>
<td>0.284</td>
</tr>
<tr>
<td>SAF</td>
<td>2.34 (0.69, 7.99)</td>
<td>0.174</td>
<td>1.50 (0.45, 5.65)</td>
<td>0.474</td>
</tr>
<tr>
<td>HRV-SDANN</td>
<td>1.69 (0.71, 4.03)</td>
<td>0.235</td>
<td>1.43 (0.59, 3.45)</td>
<td>0.423</td>
</tr>
<tr>
<td>HRV-SDNN</td>
<td>1.69 (0.71, 4.02)</td>
<td>0.239</td>
<td>1.41 (0.59, 3.40)</td>
<td>0.442</td>
</tr>
<tr>
<td>HRV-ASDNN</td>
<td>1.38 (0.56, 3.39)</td>
<td>0.478</td>
<td>1.18 (0.48, 2.91)</td>
<td>0.726</td>
</tr>
<tr>
<td>HRV-PNN50</td>
<td>0.88 (0.32, 2.38)</td>
<td>0.798</td>
<td>0.93 (0.34, 2.53)</td>
<td>0.890</td>
</tr>
<tr>
<td>HRV-HRV1</td>
<td>0.87 (0.32, 2.35)</td>
<td>0.779</td>
<td>0.72 (0.26, 1.98)</td>
<td>0.526</td>
</tr>
<tr>
<td>HRV-RMSSD</td>
<td>0.66 (0.22, 1.95)</td>
<td>0.451</td>
<td>0.68 (0.23, 2.00)</td>
<td>0.480</td>
</tr>
<tr>
<td>TWA</td>
<td>1.16 (0.45, 2.95)</td>
<td>0.763</td>
<td>0.63 (0.21, 1.88)</td>
<td>0.411</td>
</tr>
</tbody>
</table>

Figure 3-4: Kaplan-Meier curves demonstrating risk stratification of two relatively lower risk subpopulations using the upper quartile value in each population (A: TRS ≤ 4; B: TRS ≤ 4 and BNP ≤ 80pg/ml). Numbers of patients remaining in the study at each labeled time point are indicated below the respective labels.

LFHF, and DC were significantly associated with CVD (Table 3.4). After adjusting for EF and BNP, only MV and MVB remained significantly associated with CVD (HR 3.7 and 3.0, respectively, p<0.05). Figure 3-4A shows the Kaplan-Meier curves
for MVB in this population.

In a subgroup of patients with even lower CVD rate (TRS \leq 4 and BNP \leq 80pg/ml, 1-year CVD rate 1.6%), MVB and MV were significantly associated with CVD after adjusting for EF (HR 7.8 and 4.4, p<0.05). Figure 3-4B shows the Kaplan Meier curves for MVB in this population.

Similar results were found in other low-risk subpopulations based on a combination of TRS and EF and/or BNP: MVB and MV were significantly associated with CVD, but not the other ECG metrics. The HRs for MVB and MV in the validation cohort and all low risk subgroups are summarized in Table 3.2.

3.3.3 Choosing an Optimal Threshold for MVB

The results presented above used the upper quartile value as the high-risk threshold. This facilitates comparison of the various ECG metrics because equal numbers of patients were predicted to be high risk. However, such a choice is not useful clinically since it is difficult to know \textit{a priori} what the upper quartile would be for any given
set of patients. In this regard, a set cutoff value would be more useful. We therefore also evaluated a set cutoff value (2.9), corresponding to the upper quartile value in our derivation cohort, to calculate hazard ratios in our validation cohorts. Hazard ratios for MVB>2.9 are very similar to that achieved using the quartile values and are therefore not shown. We present the rates of cardiovascular deaths as a function of MVB quartiles in the validation cohort and low-risk subpopulations in Figure 3-5 demonstrating a graded increase in CVD rate with MVB in the validation cohort and all low risk subgroups.

3.3.4 Correlation of MVB with other Variables

We also measured the correlation of MVB with the other ECG metrics (Table 3.5) to ensure that MVB identified different groups of high risk patients compared to other risk metrics. The correlation coefficients were below 0.5 for all ECG metrics. In particular, MVB was uncorrelated with TWA, the other morphology-based ECG metric analyzed in this study (r=-0.044). The correlation coefficient was 0.361 for heart rate and below 0.15 for all other baseline characteristics.

3.3.5 Analysis of MVB in Specific ECG Segments

To investigate the physiology underlying the variability measured by MVB, we computed MVB using partial segments of the ECG. In the derivation cohort, the four quarters resulted in AUCs of 0.608, 0.629, 0.644, and 0.687 respectively. These AUCs were lower than that of MVB computed using the entire ECG (0.725) (Table 3.6).

3.3.6 Correlation between Time and Beat-Frequency Bands

We analyzed the inter-feature correlation in time- and beat-frequency and presented these data as a heat map (Figure 3-6). The heat map for time-frequency correlations is smooth, showing that each frequency is related to other frequencies, and this relationship changes gradually with respect to time-frequency. By contrast, the heat map for beat-frequency contains distinctive striations and square-shaped hot spots. The
Table 3.5: Correlation of MVB and MV with other risk metrics (top section) and patient factors (bottom section). To normalize for the different numerical ranges of the different metrics, continuous risk metrics were dichotomized at the upper quartile in the placebo population. For categorical risk metrics with more than two categories, the highest risk categories were used: HRT=2 and TRS≥5.

<table>
<thead>
<tr>
<th>Risk Metric</th>
<th>MVB</th>
<th>MV</th>
</tr>
</thead>
<tbody>
<tr>
<td>MV</td>
<td>0.745</td>
<td>1</td>
</tr>
<tr>
<td>HRV-SDNN</td>
<td>0.095</td>
<td>0.007</td>
</tr>
<tr>
<td>HRV-SDANN</td>
<td>0.146</td>
<td>0.078</td>
</tr>
<tr>
<td>HRV-ASDNN</td>
<td>0.083</td>
<td>0.018</td>
</tr>
<tr>
<td>HRV-RMSSD</td>
<td>-0.088</td>
<td>-0.125</td>
</tr>
<tr>
<td>HRV-PNN50</td>
<td>-0.128</td>
<td>-0.148</td>
</tr>
<tr>
<td>HRV-HRVI</td>
<td>0.139</td>
<td>0.071</td>
</tr>
<tr>
<td>HRV-LFHF</td>
<td>0.400</td>
<td>0.419</td>
</tr>
<tr>
<td>HRT</td>
<td>0.180</td>
<td>0.184</td>
</tr>
<tr>
<td>DC</td>
<td>0.382</td>
<td>0.370</td>
</tr>
<tr>
<td>SAF</td>
<td>0.204</td>
<td>0.213</td>
</tr>
<tr>
<td>TWA</td>
<td>0.044</td>
<td>0.051</td>
</tr>
<tr>
<td>BNP80</td>
<td>0.166</td>
<td>0.155</td>
</tr>
<tr>
<td>TRG</td>
<td>0.069</td>
<td>0.059</td>
</tr>
<tr>
<td>EF40</td>
<td>0.177</td>
<td>0.174</td>
</tr>
<tr>
<td>Age≥65</td>
<td>0.148</td>
<td>0.188</td>
</tr>
<tr>
<td>Gender</td>
<td>0.022</td>
<td>0.003</td>
</tr>
<tr>
<td>BMI&gt;30</td>
<td>0.054</td>
<td>0.018</td>
</tr>
<tr>
<td>Diabetes Mellitus</td>
<td>0.086</td>
<td>0.067</td>
</tr>
<tr>
<td>Hypertension</td>
<td>0.046</td>
<td>0.055</td>
</tr>
<tr>
<td>Smoker</td>
<td>-0.067</td>
<td>-0.086</td>
</tr>
<tr>
<td>Previous MI</td>
<td>0.003</td>
<td>0.016</td>
</tr>
<tr>
<td>PCI</td>
<td>0.001</td>
<td>-0.004</td>
</tr>
<tr>
<td>CHF</td>
<td>0.03</td>
<td>0.041</td>
</tr>
<tr>
<td>Ventricular Arrhythmia</td>
<td>0.058</td>
<td>0.052</td>
</tr>
<tr>
<td>Resuscitation</td>
<td>0.042</td>
<td>0.056</td>
</tr>
<tr>
<td>Creatine Clearance &lt;60</td>
<td>0.106</td>
<td>0.113</td>
</tr>
<tr>
<td>Index Event</td>
<td>0.094</td>
<td>0.068</td>
</tr>
<tr>
<td>ST depression</td>
<td>0.069</td>
<td>0.071</td>
</tr>
<tr>
<td>Prior Angiography</td>
<td>0</td>
<td>-0.001</td>
</tr>
<tr>
<td>Aspirin</td>
<td>-0.086</td>
<td>-0.068</td>
</tr>
<tr>
<td>Beta-Blocker</td>
<td>-0.042</td>
<td>-0.047</td>
</tr>
<tr>
<td>Statin</td>
<td>0.002</td>
<td>0.015</td>
</tr>
<tr>
<td>Heart Rate</td>
<td>0.361</td>
<td>0.284</td>
</tr>
</tbody>
</table>
Table 3.6: Performance of MVB using segments of the ECG

<table>
<thead>
<tr>
<th>Quarter-segment</th>
<th>AUC for MVB</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.6078</td>
</tr>
<tr>
<td>2</td>
<td>0.6293</td>
</tr>
<tr>
<td>3</td>
<td>0.6443</td>
</tr>
<tr>
<td>4</td>
<td>0.6868</td>
</tr>
<tr>
<td>Entire Segment</td>
<td>0.7253</td>
</tr>
</tbody>
</table>

Figure 3-6: Correlation between time-frequency bands (left) and beat-frequency bands (right).

Most obvious striations are at 0.1 and 0.5 cycles/beat, or every 10 and every 2 beats. Further, square hot spots can be observed at 0.13 to 0.17 cycles/beat (every 5 to 7 beats). We also note that the beat-frequency band at every 2 beats is uncorrelated with all other bands.

3.4 Discussion

3.4.1 Predictive Performance

We evaluated the ability of several ECG-based risk metrics to risk stratify patients in conjunction with well-established clinical risk tools such as the TRS, EF, and BNP. Existing risk metrics typically only identify a subset of the high-risk patients. For example, patients with a high TRS were at the highest relative risk of CVD, but only accounted for half of all CVD. Therefore, a low-to-moderate risk population
still contains a significant number of adverse events, and the lack of appropriate risk stratification in such populations remains a clinical deficiency. To our knowledge, this is the first study to systematically evaluate the performance of these ECG metrics on patients who would not be identified as high risk based on multiple risk measures.

We find that in a population with a low-to-moderate TRS, MV and MVB were the only ECG metrics that were significantly associated with CVD after adjusting for EF and BNP. This relationship was similar in other low-risk subpopulations. Our results in the lower risk subpopulations also reveal an interesting trend. MVB has a higher hazard ratio when evaluated in populations with lower CVD rates, suggesting that MVB is particularly good at identifying high risk patients that are missed by other risk measures. This trend is also present, but to a smaller extent, for MV. Unfortunately, the small numbers of adverse outcomes in these populations makes it difficult to assess the statistical significance of the improvement of MVB relative to MV.

However, we are still optimistic that MVB represents an improvement relative to MV for several reasons. First, it is easier to interpret the diagnostic beat-frequency compared to the diagnostic frequency. This is because the MD time series quantifies differences between the morphology of adjacent heartbeats. Therefore, each MD value is only defined at each heartbeat, and events cannot occur in between. Thus it is more natural to characterize events with respect to heartbeats for this signal. Second, the inter-frequency correlations in time- and beat-frequency are markedly different. The most striking differences are the square boxes of high correlation and the horizontal and vertical striations at 0.10 and 0.50 cycles/beat. The square boxes demarcate beat-frequency bands that are highly correlated with each other, enabling easy selection of groups of bands for physiological studies. By contrast, the highly correlated groups of time-frequency bands are not as clearly demarcated.

3.4.2 Physiological Basis

The physiological basis of MV and MVB is also of interest. The MD time series measures beat-to-beat morphologic changes in an ECG, which can arise when unstable
islands of ischemia throughout the myocardium present in a probabilistic or random manner [16, 95]. Because high frequencies of abnormal events are cause for greater concern in general, it is not surprising that the repeating frequency of MVB (diagnostic beat-frequency) corresponds to high beat-frequencies, or equivalently, low values of “every x beats.”

While ischemia often manifests during myocardial repolarization, it can also result in morphologic changes elsewhere in the cardiac cycle; e.g., in the PR segment and QRS complex. Therefore we strove to develop a metric that would quantify the morphologic variability using the entire cardiac cycle. Consistent with this hypothesis, MVB variants that quantify morphologic differences in partial segments of the ECG segment result in poorer discriminative performance.

In this respect, one may see MV and MVB as a more general, data driven version of TWA. Instead of focusing on just the ST segment and T wave, we measure variability in all segments of the ECG. Instead of measuring alternating changes, we measure differences in a diagnostic frequency that is optimized using patient data. However, TWA measures a repeating ABAB pattern, which would produce consistently high values in the MD time series. This consistently high average value would correspond to the DC (direct current, frequency = 0) component of the frequency spectra, which is not used in either MV or MVB. As such, MV and MVB measure distinct ECG activity that is not captured by TWA. This is reflected in the low correlation between MV and MVB with TWA.

3.5 Limitations

Our study has limitations. First, EF and BNP measurements were not available for all patients, reducing the population size relative to the original patient population. Our goal of identifying high risk patients that current risk metrics miss further reduces the pool of patients available for validation. This made it challenging to assess the statistical differences between MVB and MV.

Furthermore, we used a very simple approach of summing energies in beat-frequency
bands. One may expect that applying different weights to different frequency bands may result in better performance. We tested this as a proof of concept by first merging the derivation and validation cohorts to obtain a dataset of reasonable size for machine learning purposes. We defined the outcome to be cardiovascular death within 90 days, the shorter of the follow-up periods in the two datasets. Next, we extracted 50 features for each patient, corresponding to 50 beat-frequency bands from 0.01 to 0.50 cycles/beat. We then trained a linear support vector machine on a randomly selected subset of the merged dataset and tested it on the remainder. We obtained good performance (AUC=0.78 in the test set, compared to 0.73 for MV). Encouragingly, our method also found that the most important features were 0.01, 0.02, 0.17, and 0.50 cycles/beat, or correspondingly, every 100, 50, 5.9 and 2 beats. The fact that the most predictive beat-frequencies were exactly at or close to integral values of “every x beats” despite being given all 50 beat-frequency bands lends further credence to beat-frequency analysis.

However, despite the substantially improved performance, it was difficult to assess the significance of the improvement for several reasons. First, the performance of MV on this test set may be overestimated because it was originally optimized on the derivation cohort, which comprised a subset of the new test set. Second, TRS, EF, and BNP were not available for the derivation cohort, preventing assessment of the additional clinical utility of this approach relative to the clinical variables. Lastly, the small size of the test set prevented meaningful subgroup analysis for patients who did have the clinical measurements. Therefore, we presented only methods and results for the simplified version, MVB, which only computes beat-frequencies corresponding to integral values of “every x beats,” and was trained/tested on the same datasets as the original MV.

### 3.6 Future Work

A straightforward technical extension to this work would be to train a machine learning model, as outlined in the previous section, specifically on patients who are low risk
as determined by TRS, EF, and BNP. This is challenging because by definition, few adverse events occur in low risk populations. Therefore a larger patient population is required to obtain adequate numbers of events.

Clinically, MVB and MB could be evaluated on additional patient populations to validate the risk estimates and determine optimal treatment protocols. For example, MV was depressed in patients who were prescribed ranolazine [96], a drug that may have anti-arrhythmic effects [85]. This suggests that MV and MVB could be used to quantify the effects of certain drugs. It would be interesting to see if certain therapies were more (or less) effective in the patient populations identified by MVB and MV.

3.7 Conclusion

We have shown that an ECG morphology-based metric may provide incremental risk stratification in patients who would normally be considered to be “low-to-moderate” risk on presentation as determined by a clinical risk score, and levels of BNP and EF. To our knowledge, this is the first study to evaluate the performance of multiple ECG metrics in low-to-moderate risk patients with comprehensive comparison to both BNP and EF, and therefore may further improve development of risk stratification methods (both morphology-based and otherwise) specifically for low-risk populations, where a significant number of cardiovascular complications continue to occur.
Chapter 4

Risk Stratification Using ECG: Heart Rate Variability

4.1 Introduction

Chapter 3 used beat-frequency to improve the risk metric Morphologic Variability, which quantifies beat-to-beat variability in ECG morphology in the frequency domain. In this chapter, we apply beat-frequency to the study of heart rate variability (HRV) [70].

Frequency domain HRV measures have been shown to patients who are at increased risk of adverse medical outcomes in multiple studies [21, 20, 50, 49, 98, 90]. As recommended by the HRV Task Force [70], most studies that analyze frequency domain HRV do so in the time-frequency domain. While some authors have investigated the beat-frequency domain [57, 73, 72, 69, 86, 97, 103], there have been no published studies of the utility of beat-frequency in predicting adverse outcomes.

In this study, we tested the hypothesis that beat-frequency applied to an established HRV measure, Low Frequency / High Frequency (LF/HF, a ratio of energy in two bands) [70] will improve prediction performance for cardiovascular death post non-ST elevation acute coronary syndrome (NSTEACS). We focused on LF/HF because LF/HF was the best performing HRV measure on our datasets [58, 93]. However, the pre-defined time-frequency LF and HF bands were not explicitly selected for
optimal risk stratification. Instead, the bands were defined in two large clinical studies [21, 20], and were standardized in a seminal review [70]. Thus, we used machine learning to discover the time- and beat-frequency bands most useful in predicting cardiovascular death post NSTEACS.

### 4.2 Methods

#### 4.2.1 Data and Outcomes

The work described in this chapter used three datasets (D1, D2, and D3), as summarized in Table 4.1. These datasets were derived from two clinical trials [67, 25], and contains ECG recordings of patients after NSTEACS.

D1 and D2 each contained about 2,300 patients from the placebo and treatment arms of the first clinical trial [67]. There were 93 and 77 cardiovascular deaths (CVD) in D1 and D2 respectively within the median follow-up of one year. Most of our results were focused on D1 because D1 was the largest. In addition, patients in D2 were prescribed ranolazine, a drug that appears to have anti-arrhythmic properties [85] and thus affect ECG measures. Thus, we reserved D2 as an additional validation.

Table 4.1: Datasets and Patient Characteristics. CVD = Cardiovascular Death, IQR = Interquartile Range (the values at the 25<sup>th</sup> and 75<sup>th</sup> percentiles, MI = Myocardial Infarction.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical Trial</td>
<td>MERLIN placebo</td>
<td>MERLIN treatment</td>
<td>DISPERSE-2</td>
</tr>
<tr>
<td>N</td>
<td>2302</td>
<td>2255</td>
<td>765</td>
</tr>
<tr>
<td># of CVD (%)</td>
<td>93 (4.0)</td>
<td>77 (3.4)</td>
<td>14 (1.8)</td>
</tr>
<tr>
<td>Median follow-up</td>
<td>1 year</td>
<td>1 year</td>
<td>90 days</td>
</tr>
<tr>
<td>Age, median (IQR)</td>
<td>64 (55-72)</td>
<td>63 (55-72)</td>
<td>62 (54-71)</td>
</tr>
<tr>
<td>Age ≥75 (%)</td>
<td>17</td>
<td>17</td>
<td>14</td>
</tr>
<tr>
<td>Female (%)</td>
<td>35</td>
<td>34</td>
<td>36</td>
</tr>
<tr>
<td>Diabetes mellitus (%)</td>
<td>34</td>
<td>33</td>
<td>24</td>
</tr>
<tr>
<td>Hypertension (%)</td>
<td>74</td>
<td>73</td>
<td>69</td>
</tr>
<tr>
<td>Current smoker (%)</td>
<td>25</td>
<td>27</td>
<td>56</td>
</tr>
<tr>
<td>Previous MI (%)</td>
<td>33</td>
<td>33</td>
<td>25</td>
</tr>
</tbody>
</table>
dataset for our machine learning models, but did not merge it with D1.

D3 consisted of 765 patients from a second clinical trial [25], and contained 14 CVD within 90 days. We did not split D3 into separate placebo and treatment datasets because, unlike dataset D2, patients in this trial were not prescribed an experimental drug that affects the ECG. This dataset was not merged with D1 because of the shorter duration of follow-up. We used D1 as an additional validation dataset for our machine learning model.

For all three datasets, up to 7 days of ambulatory ECG signals recorded at 128 Hz were available for each patient. We used the first 24 hours from each patient.

The reader might notice that these datasets are similar to those used in Chapter 3. Indeed, the derivation cohort for MVB was D3, and the validation cohort was a subset of D1. We selected the datasets for MVB based on the datasets used in the original MV metric, and the desire to evaluate the risk metrics relative to clinical risk factors. In this chapter, we are not building on another risk metric developed on one of these datasets. Therefore, we used the largest dataset (D1) for machine learning.

### 4.2.2 Time-frequency LF/HF

We computed the time-frequency LF/HF using the established LF (0.04 to 0.15 Hz) and HF (0.15 to 0.40 Hz) bands [70]. We first segmented the 24-hour long ECG signal to find individual heartbeats, and used the Signal Quality Index [54] to extract the normal beats. These steps are described in detail in [93], and are identical to the procedure used in Chapter 3.

Next, we divided the heart rate time series into five-minute intervals and estimated the power spectral density in the time-frequency domain. There are many ways to compute the frequency domain power spectrum; as in Chapter 3, we use the Lomb-Scargle periodogram, which provides a natural way to handle unevenly sampled data [60 28]. For the heart rate time series, this uneven sampling has two sources: the inherent heart rate variations in the underlying ECG, and the removal of ectopic beats or artifacts. We then computed the ratio of the energy in the LF band to the HF band for each five-minute window, and defined the median value of all these
LF/HF ratios to be the LF/HF for that patient. A low value of LF/HF is known to be associated with poorer prognosis [21, 20, 88].

4.2.3 Beat-frequency LF/HF

Beat-frequency LF/HF was also obtained from the power spectrum of the heart rate time series, however, the heartbeat indices were used as the “temporal” reference. As in Chapter 3, where beats or parts of the signal were removed in the preprocessing step, the number of removed beats was estimated using the average time intervals of the heartbeats immediately adjacent to the gap.

Diagnostic bands for beat-frequency LF/HF were based on a prior study that applied LF and HF beat-frequency bands to normal adults as well as those with coronary artery disease and congestive heart failure: 0.03 to 0.14 (LF band) and 0.14 to 0.40 cycles/beat (HF band) [103]. The remaining steps in beat-frequency LF/HF computation were identical to the time-frequency LF/HF computation.

4.2.4 Machine Learning

The LF and HF time- and beat-frequency bands above were not optimized for prediction of adverse events, but were instead based on previous physiological studies and observations of where peaks are located in the power spectra. In this section, we present a machine learning procedure to learn the predictive time- and beat-frequency bands as well as the relative importance (weights) of each band. We term this approach Weighted HRV (WHRV).

Figure 4-1 summarizes and compares the LF/HF and WHRV approaches. The initial steps were similar: we obtained a heart rate time series, divided it into five-minute windows, and converted each window into each of the two frequency domains. In the WHRV approach, we next measured the energy in 50 bands: 0.01 to 0.50 Hz for time-frequency and 0.01 to 0.50 cycles/beat for beat-frequency. These 50 values formed a feature vector for each patient.

The feature vectors were used as input to an $L_1$-regularized logistic regression
Figure 4-1: Comparison of HRV metrics assessed in this study. In blue: time- and beat-frequency versions of LF/HF, the ratio of energy in two pre-defined frequency bands. In red: time- and beat-frequency versions of the energy in bands weighted using machine learning.

Algorithm, which selects the most important features by optimizing:

\[
\hat{w} = \arg \min_w \sum_j |w_j| + C_+ \sum_{i \in +} \log(1 + e^{-y_i w^T x_i}) + C_- \sum_{i \in -} \log(1 + e^{-y_i w^T x_i})
\]

where \( w \) is the weight vector that quantifies the importance of each feature — \( i \) indexes patients, \( j \) indexes features, \( - \) and \( + \) indicate the groups of patients that survived and died respectively, \( y \) is the label for each patient (\(-1\) if patient survived, \(+1\) if patient died), and \( x \) is the feature vector for each patient. \( C \) is a cost parameter that balances the regularization piece (leftmost summation) and the loss function (other
two summations). \( C_+ \) and \( C_- \) are the cost parameters for the patients who died and survived at the end of one year, respectively. We apply a different (or asymmetric) cost parameter for these two classes of patients because only 4% of patients died, leading to a high class imbalance. We set \( C_+ = k \times C_- \), where \( k \) is the class imbalance ratio.

For each value of \( C_- \), the algorithm outputs a weight vector \( \mathbf{w} \). We optimized \( C_- \) in the range \( 10^c, c = -6, -5, -4, ..., 2 \), and selected the \( C_- \) with the highest average cross validation score. The evaluation score will be described in the next section. In our experiments, we used 5 repeats of 5-fold cross validation to find the optimal \( C_- \). Finally, we trained a single model using the optimized \( C_- \) and the training set, and evaluated the model on the test set.

For each patient, the WHRV was defined as the output of the scalar valued dot product of the feature vector with the weight vector, or \( \mathbf{w}^T \mathbf{x}_i \). Higher values indicate higher risk.

4.2.5 Evaluation of Machine Learning Models

To develop and test WHRV, we first randomly split the patients in dataset D1 into a training set and a test set in a 2:1 ratio. This training/test split procedure ensures that models are evaluated based on their performance on patients that the models have previously not seen. We also ensured that both training and test sets contained the same rates of deaths using a stratified sampling approach: we performed the 2:1 split separately for each class (survived and died).

We evaluated our a machine learning model using the area under the receiver operating characteristic curve (AUC) \([4]\), for the outcome of cardiovascular death within 1 year. The AUC is described in Section 2.1.3, but briefly, the AUC measures prediction performance and ranges from 0 to 1. 0.5 indicates random performance, and higher values are better.

We repeated this training/test split 1,000 times to reduce the effects of selecting an overly optimistic or pessimistic test set. In other words, there are 1,000 training sets and 1,000 test sets. We report results averaged over the 1,000 test sets. To
evaluate the performance of these models in the additional holdout datasets D2 and D3, we computed the AUC for each of the 1,000 models in each holdout dataset, and report the average AUC.

The relative importance of the 50 time- and beat-frequency features were analyzed across the 1,000 different models by dividing the weight of each feature by the maximum absolute weight in each model. After this procedure, the largest absolute value of the normalized weights in each model will be 1. We estimated the variation in feature weights by first computing the standard deviation of each normalized feature weight across all 1,000 models to obtain a standard deviation for each feature. We summarized these 50 standard deviations by taking the average. A higher value indicates a higher variation of feature weights across different training sets.

4.2.6 Comparison with Clinical Measures

We evaluated the performance of our model relative to several clinical variables, the TIMI Risk Score (TRS) [9], left ventricular ejection fraction (EF), and B-type natriuretic peptide (BNP) [32]. We used the same high-risk cutoffs as in Chapter 3: TRS \( \geq 5 \), EF \( \leq 40\% \), and BNP > 80 pg/ml. Because EF and BNP were only measured in 47% of the patients [58], we build separate multivariable models: a model that includes only TRS and uses all the patients, and a model that includes TRS, EF, and BNP, and uses only the patients with measured values of both EF and BNP.

4.2.7 Statistical analysis

In the LF/HF section, we assessed the statistical significance of differences between the AUCs of two metrics (time- and beat-frequency versions) on the entire dataset [33]. In the machine learning section, we assessed the difference of AUCs across 1,000 test sets using a two-tailed paired t-test [82].

In addition, we computed the hazard ratio (HR) of the different HRV risk metrics using the Cox proportional hazards regression model [30]. We dichotomized the continuous machine learning predictions at the upper quartile of each test set; thus
the HR indicates the hazard ratio of the upper quartile compared to the remaining patients. We adjusted the HR for TRS, EF, and BNP using binary variables defined by the high-risk cutoffs described in the previous section. Thus, we present three HRs: unadjusted, adjusted for TRS, and adjusted for all of TRS, EF, and BNP.

4.2.8 Correlation Between Time- and Beat-Frequency

To ensure that the two frequency domains were not trivially different, we quantified the correlation coefficient between the two types of frequency bands. For example, if a given time-frequency band in the range \( x_1 \) to \( x_2 \) Hz is completely identical in all patients to some beat-frequency band in the range \( y_1 \) to \( y_2 \) cycles/beat, then any differences in prediction performance could be eliminated by selecting a better band.

Closest Beat-frequency to Established Time-frequency Bands

First, we searched for the closest beat-frequency band to the established time-frequency bands. To ensure that patients are represented equally, we randomly selected 10 five-minute windows from all 2302 patients in dataset D1. Next, we computed the energy in the time-frequency LF band and the entire beat-frequency spectra from 0.01 to 0.50 cycle/beat for these patients’ five-minute windows. Then, we computed the correlation coefficient between the LF band and all possible beat-frequency bands, and selected the beat-frequency band with the highest correlation. We repeated this procedure for the time-frequency HF band.

Closest Time-frequency to Learned Beat-frequency Bands

First, we searched for the closest time-frequency to the low and high beat-frequency bands learned from our machine learning approach. We used the learned bands here because we were able to learn distinct prediction bands in beat-frequency (but not time-frequency). Because the location of time-frequency bands may depend on heart rate, we stratified our analysis by heart rate. First, we selected all five-minute windows that had an average heart rate of 50±1 beats per minute. Next, we computed the
low beat-frequency band and entire time-frequency band from 0.01 to 0.50 Hz for each five-minute window. Lastly, we computed the correlation coefficient between the low beat-frequency band and all possible time-frequency bands, and selected the time-frequency band with the highest correlation. We repeated this analysis for five-minute windows with an average heart rate of 60±1, 70±1, 80±1, 90±1, and 100±1. Finally, we repeated all the steps for the high beat-frequency band.

4.3 Results

The results section is arranged as follows. We first compare time- and beat-frequency versions of LF/HF in the entire dataset D1. Next, we compare time-and beat-frequency machine learning models (WHRV) on test sets sampled from D1 and additional holdout sets D2 and D3. We then report the predictive time- and beat-frequency bands found by machine learning. Finally, we compare the machine learning models with LF/HF computed on the same test sets.

4.3.1 LF/HF in Time- and Beat-frequency

The AUC for LF/HF in time- and beat-frequency in dataset D1 were significantly different, 0.713 for time-frequency and 0.731 for beat-frequency, p<0.01. The quartile cutoffs for time and beat frequency were 0.841 and 0.832 respectively, resulting in hazard ratios (HR) of 3.7 and 4.6. After adjusting for TRS, the HRs were 3.0 and 3.9 respectively. After adjusting for TRS, EF, and BNP, the HR were both 2.4. All HRs were significant (p<0.01).

4.3.2 Machine Learning (WHRV) in Time- and Beat-frequency

The AUC for beat-frequency WHRV was significantly higher than that of time-frequency in each of the test sets (Table 4.2 0.753 versus 0.704, p<0.001). Similar improvements were observed in the two additional holdout sets (Table 4.3 0.728 versus 0.691, and 0.738 versus 0.679, p<0.001 for both improvements).
Table 4.2: Area Under Curve (AUC) of time- and beat-frequency HRV measures averaged over 1,000 test sets. Bold indicates the higher AUC in each row, and * indicates the highest AUC.

<table>
<thead>
<tr>
<th>Risk Measure</th>
<th>AUC (std error)</th>
<th>Beat</th>
<th>Time</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>LF/HF (Pre-defined bands)</td>
<td>0.7303 (0.0011)</td>
<td>0.7042 (0.0012)</td>
<td>&lt;0.001</td>
<td></td>
</tr>
<tr>
<td>WHRV (Machine learning)</td>
<td>0.7526 (0.0012)*</td>
<td>0.7036 (0.0012)</td>
<td>&lt;0.001</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.3: AUC of WHRV models evaluated on holdout sets D2 and D3.

<table>
<thead>
<tr>
<th>Dataset, N (# of deaths)</th>
<th>Weighted HRV AUC (std error)</th>
<th>Beat</th>
<th>Time</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>D2 (Holdout Set 1), 2255 (77)</td>
<td>0.7277 (0.0002)</td>
<td>0.6905 (0.0004)</td>
<td>&lt;0.001</td>
<td></td>
</tr>
<tr>
<td>D3 (Holdout Set 2), 765 (14)</td>
<td>0.7384 (0.0003)</td>
<td>0.6793 (0.0007)</td>
<td>&lt;0.001</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.4: Hazard Ratio (HR) averaged over 1,000 test sets. 95% confidence intervals (CI) reported in parenthesis. Bold indicates the highest HR in each row (p<0.001 for unadjusted and adjusted for TRS; in the last row only WHRV in beat-frequency has a CI that does not include 1). *: In 7 out of 1,000 test sets, less than 30% of the patients had measured values of both EF and BNP, and therefore these test sets were excluded.

<table>
<thead>
<tr>
<th></th>
<th>WHRV HR (95% CI)</th>
<th>LF/HF HR (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Beat</td>
<td>Time</td>
</tr>
<tr>
<td>Unadjusted</td>
<td>5.19 (2.60,9.45)</td>
<td>3.70 (1.98,6.76)</td>
</tr>
<tr>
<td>Adj. for TRS</td>
<td>4.38 (2.23,8.62)</td>
<td>3.13 (1.72,5.89)</td>
</tr>
<tr>
<td>Adj. for TRS,EF,BNP*</td>
<td>2.91 (1.05,8.68)</td>
<td>1.99 (0.69,5.73)</td>
</tr>
</tbody>
</table>
Figure 4-2: Boxplot of normalized weights of time- and beat-frequency machine learning models trained on the same 1,000 training splits. Quartiles are represented by the edges of lines, box, and central dot, while circles indicate outliers. The average standard deviations of the weights are 0.057 in beat-frequency and 0.153 in time-frequency.

In our machine learning experiments, each of the 1,000 randomly chosen training sets generated a distinct weight vector. We observed a greater consistency in beat-frequency features compared to time-frequency in these 1,000 weight vectors: the average standard deviation of the normalized beat-frequency feature weights is less than half that of the time-frequency value (Figure 4-2, 0.057 versus 0.153, p<0.001). Moreover, beat-frequencies between 0.03 and 0.07 cycles/beat have consistently negative weights across the 1,000 different models — suggesting that high power in these frequency bands is associated with decreased rate of one-year cardiovascular death. By contrast, beat-frequencies between 0.18 and 0.24 cycles/beat are associated with positive weights — suggesting that high power in these frequency bands is associated with an increased rate of one-year cardiovascular death. Since the models built using time-frequency features show much more variation across the 1,000 different models, it is difficult to make reliable inferences about the relative importance of different time-frequency bands.

We present the HR’s of the models before and after adjusting for clinical variables.
Figure 4-3: Positive and negative predictive values of WHRV using different high risk thresholds.

The unadjusted HR’s were 3.7 and 5.2 for time- and beat-frequency respectively. In multivariable analysis, the HR’s were 3.1 and 4.4 after adjusting for TRS, and 2.0 and 2.9 after adjusting for TRS, EF and BNP. The estimated confidence intervals did not include 1 for all values except time-frequency adjusted for all three clinical variables.

We evaluated the additional value of using machine learning to weight frequency bands (WHRV) compared to using a ratio of pre-defined bands in the LF/HF approach. We computed the AUC for the time- and beat-frequency versions of LF/HF on the same test sets as the machine learning models (Table 4.2). Machine learning left the AUC of time-frequency LF/HF unchanged. However, machine learning substantially increased the AUC of beat-frequency LF/HF (0.730 to 0.753, p<0.001). Similar results were obtained with respect to the HR; the adjusted HR using machine learning was higher than those using pre-defined bands for beat-frequency but not time-frequency (Table 4.4).

The results above used the quartile as the “high-risk” cutoff. To explore the utility of using other cutoffs, we compared the positive predictive values and negative predictive values of these four metrics over all possible cutoffs from the 5th to the 95th percentile, in increments of 10 (Figure 4-3). Beat-frequency WHRV demonstrated the highest positive predictive value from the 25th to the 85th percentile, demonstrating robustness to the choice of the cutoff. The negative predictive values of all the
Table 4.5: Final machine learning model in beat-frequency. Only features that are selected by the machine learning algorithm are shown. To compute the WHRV risk metric for a new patient, each feature is converted to a z-score by subtracting the mean and dividing by the standard deviation. The final risk metric is the sum of the products of the z-scores and the corresponding weights.

<table>
<thead>
<tr>
<th>Feature (cycles/beat)</th>
<th>0.03</th>
<th>0.05</th>
<th>0.06</th>
<th>0.07</th>
<th>0.19</th>
<th>0.22</th>
<th>0.23</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>8.413</td>
<td>4.124</td>
<td>3.148</td>
<td>2.461</td>
<td>0.875</td>
<td>0.941</td>
<td>0.952</td>
<td>1.546</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>3.656</td>
<td>1.647</td>
<td>1.306</td>
<td>1.105</td>
<td>0.6</td>
<td>0.737</td>
<td>0.622</td>
<td>5.9</td>
</tr>
<tr>
<td>Weight</td>
<td>-0.1682</td>
<td>-0.4518</td>
<td>-0.1435</td>
<td>-0.0032</td>
<td>0.0638</td>
<td>0.1791</td>
<td>0.016</td>
<td>0.0144</td>
</tr>
</tbody>
</table>

measures were above 96% at all cutoffs, suggesting utility in identifying patients who are not at high-risk.

Another factor in evaluating a risk model is how well its predicted probability of adverse event matches the actual probability of events. This is called the calibration of the model. Because of the asymmetric cost parameter in our logistic regression model, the predicted probabilities are not calibrated by default. However, the positive predictive value plot in (Figure 4-3) allows estimation of the probability of death in high-risk patients based on any given cutoff for our risk model. For example, high-risk patients defined using a 75th percentile cutoff had an approximate death rate of 10% in the following year.

Finally, we trained a WHRV model using the entire dataset for use in future patient populations (Table 4.5). These parameters could be used in future studies of similar patient populations without having re-training.

### 4.3.3 Correlation between Time- and Beat-frequency

Finally, we quantified the similarity between time- and beat-frequency to ensure that the two frequency domains were not rescaled versions of each other. The most similar beat-frequency bands to the LF and HF time-frequency bands were 0.04-0.13 cycles/beat and 0.10-0.37 cycles/beat, with correlation coefficients of 0.63 and 0.81 respectively. These correspond to 40% and 66% of the explained variance, respectively.

The most similar time-frequency bands in to our learned beat-frequency bands
Table 4.6: Most highly correlated time-frequency bands to the two learned beat-frequency bands. The time-frequency bands show a trend towards higher frequencies at higher heart rates, and the explained variances are almost always less than 67%.

<table>
<thead>
<tr>
<th>Heart Rate [beats per min]</th>
<th>Time-frequency [Hz]</th>
<th>Correlation [r]</th>
<th>% variance explained [R²]</th>
<th>Time-frequency [Hz]</th>
<th>Correlation [r]</th>
<th>% variance explained [R²]</th>
</tr>
</thead>
<tbody>
<tr>
<td>50±1</td>
<td>0.04 - 0.06</td>
<td>0.457</td>
<td>21</td>
<td>0.15 - 0.30</td>
<td>0.775</td>
<td>67</td>
</tr>
<tr>
<td>60±1</td>
<td>0.03 - 0.07</td>
<td>0.933</td>
<td>87</td>
<td>0.18 - 0.24</td>
<td>0.794</td>
<td>65</td>
</tr>
<tr>
<td>70±1</td>
<td>0.04 - 0.10</td>
<td>0.514</td>
<td>33</td>
<td>0.21 - 0.28</td>
<td>0.762</td>
<td>63</td>
</tr>
<tr>
<td>80±1</td>
<td>0.04 - 0.10</td>
<td>0.759</td>
<td>58</td>
<td>0.24 - 0.32</td>
<td>0.814</td>
<td>66</td>
</tr>
<tr>
<td>90±1</td>
<td>0.05 - 0.12</td>
<td>0.73</td>
<td>53</td>
<td>0.27 - 0.37</td>
<td>0.803</td>
<td>64</td>
</tr>
<tr>
<td>100±1</td>
<td>0.05 - 0.12</td>
<td>0.789</td>
<td>62</td>
<td>0.30 - 0.40</td>
<td>0.812</td>
<td>66</td>
</tr>
</tbody>
</table>

are shown in Table 4.6. With increasing heart rate, the most similar time-frequency bands shifted towards higher Hz. This trend is most obvious for the higher beat-frequency band, 0.18 to 0.24 cycles/beat. At 50 beats per minute (bpm), the most similar time-frequency band was 0.15 to 0.20 Hz. At double the heart rate (100 bpm), the most similar time-frequency band doubled, to 0.30 to 0.40 Hz. Thus a third of the variance remained unexplained between time- and beat-frequency. We cannot measure one frequency domain using appropriately chosen bands in the other.

4.4 Discussion

Since the autonomic nervous system is an important regulator of heart rate, analysis of HRV can provide useful information about autonomic tone. Quantitative metrics like HRV that provide insight into the autonomic tone are typically calculated using time-frequency. This is justified by the fact that changes in autonomic tone vary relatively slowly with respect to the heart rate, and are therefore believed to not be a function of the precise timing of cardiac beats [18, 13]. Time-frequency domain HRV metrics have been shown to provide useful prognostic information after a myocardial infarction in multiple studies [21, 20, 50, 49, 98, 90] and after NSTEACS [58, 93]. However, though time-frequency HRV measures are associated with elevated risk after NSTEACS, these risk measures miss significant numbers of deaths [58].
4.4.1 Physiological Basis of Beat-Frequency Improvements

Prior studies have found that beat-frequency improves intra-subject and inter-subject consistency in the location of characteristic frequency bands in humans [57], and across species (humans, rats, and dogs) [83]. Our results suggest that quantifying LF/HF in terms of beat-frequency instead of time-frequency improves the ability to identify high-risk patients after a NSTEACS.

Analysis of the relationship between the time-frequency spectrum and the average heart rate provides insight into the improved performance. In one study, young male subjects exercised on a cycle ergometer at pre-specified exercise intensities [76]. Computation of the time-frequency HRV spectra revealed that the center frequency of the HF band (but not the other bands) increased with increasing exercise intensity. Figure 4-4 shows that as a function of heart rate, the center frequencies of the HF band increased from 0.24 to 0.48 Hz. Because the heart rate may vary considerably over long ECG recordings, the peaks that LF/HF is intended to measure may shift away from the pre-specified time-frequency bands [19]. By contrast, an estimate of the center frequency of the corresponding HF band in beat-frequency varies from 0.17 to 0.19 cycles/beat over the same range of heart rates (Figure 4-4). Because beat-
frequency bands are less sensitive to changes in average heart rate, they allow patients and ECG segments with different average heart rates to be directly compared.

In our data, among ECG signals with the same average heart rate, a third of the variance between the most similar time- and beat-frequency bands remains unexplained (Table 4.6). This indicates that the differences between time and beat-frequency go deeper than mere average heart rate. Indeed, it is known that the location of the HF band is influenced by the respiratory rate [23, 77]. Thus including information about the respiratory rate may further improve predictive performance of time-frequency metrics [10]. Unfortunately, the difficulty of reliably extracting the respiratory rate from Holter ECG signals prevents us from testing this hypothesis, and such data are not routinely obtained along with ECG tracings. By contrast, the computation of beat-frequency does not require the respiratory rate and can therefore be retrospectively applied to existing HRV datasets and also applied to new HRV studies without additional equipment.

4.4.2 Asymmetric Benefits of Machine Learning

The fact that beat-frequency WHRV (using machine learning) outperformed beat-frequency LF/HF (using pre-defined bands) highlights the benefit of adopting a data-driven, machine learning approach to analyzing heart rate spectra. An analysis of the weights of the beat-frequency bands finds trends that are similar to what has been observed with trends in standard time-frequency LF/HF measurements. In particular, the signs of the beat-frequency bands weights (negative at 0.03 to 0.07 cycles/beat and positive at 0.18 to 0.24 cycles/beat) are consistent with the notion that lower values of the LF/HF are associated with adverse events [88]. Our data also suggest that using features derived from the beat-frequency spectrum yields logistic regression models that have improved performance relative to logistic regression models derived from the time-frequency spectrum.

Applying machine learning to time-frequency bands does not yield a significant improvement in either the AUC or the unadjusted or adjusted hazard ratios relative to the pre-defined time-frequency bands. That machine learning improves risk stratifica-
tion with beat-frequency bands but not time-frequency bands can be explained by the
greater consistency of feature weights in beat-compared to time-frequency. Because
differences in average heart rate are associated with shifts in the time-frequency do-
main spectra, the predictive time-frequencies may differ between patients and across
time for each patient. Thus time-frequency features may not generalize across pa-
tients or even between segments of different heart rates for the same patient. This
leads to unchanged or even decreased performance despite an attempt to learn the
predictive time-frequencies.

Finally, several relative performance differences were changed by adjusting for
other risk factors adjusting. For example, beat-frequency LF/HF had a higher AUC
and TRS-adjusted HR relative to time-frequency. However, both versions had the
same HR after further adjusting for EF and BNP. This indicates that the additional
high risk patients identified by beat-frequency were also identified by EF and BNP.
Such differences could be addressed by building risk models specifically for the patients
missed by EF and BNP, as suggested in Chapter 3.

4.5 Limitations

In our feature engineering process, we divided each day-long ECG signal into five-
minute windows, and defined each feature to be the 90th percentile of the energies
in that frequency across all five-minute windows. As a consequence, each feature
may be extracted from different five-minute windows across the day. If a patient
is determined to be high risk and has high feature values corresponding to several
different frequencies, it is difficult to identify a unique five-minute ECG signal that
was responsible for the high feature value, and thus, high risk.

Our approach also does not leverage the knowledge that nearby frequencies are
correlated. For example, the power at 0.01 and 0.02 (Hz or cycles/beat) may be
correlated both because of physiological reasons and technical reasons (e.g., spectral
leakage). One potential solution is to clustering frequencies into wider bands (e.g.,
0.01 to 0.04 cycles/beat as a single feature), and another solution may be to use
structured sparsity methods such as group lasso \cite{106} to inform the model to select
groups of adjacent features.

\section{4.6 Future Work}

An important extension to this work would be physiological studies of HRV to derive a
better understanding of the relationship between respiratory rate and beat-frequency.
For example, heart rate data collected from patients breathing at controlled, pre-
specified rates \cite{23} could be analyzed in beat-frequency. Based on our work, we
expect substantially less variation in the location of peaks in beat-frequency relative
to time-frequency. However, such an analysis would provide a quantitative estimate
of the variation in beat-frequency at different respiratory rates. Such a study could
then be extended from the laboratory-controlled conditions to ambulatory, everyday
recordings of both respiratory rate and ECG. Finally, the analysis could be performed
in patient populations such as those after a heart attack to determine if respiratory
rate could provide additional useful information in the context of risk stratification.
For example, could we use the respiratory rate to warp the time-frequency spectra so
that the correct HF peak could be captured, and how would this compare in terms
of risk stratification performance with beat-frequency?

In addition, our methods could be used to identify groups of patients for targeted
therapy, similar to future work proposed in Chapter \ref{chap:future_work}. Our machine learning approach
has also produced two frequency bands that can be used to define a new LF/HF in
beat-frequency, without explicitly weighting the bands using our learned model.

\section{4.7 Conclusion}

Applying beat-frequency and machine learning to heart rate variability metrics im-
proves our ability to identify patients at elevated risk of cardiovascular death post
NSTEACS. Our machine learning approach in beat-frequency also reveals predictive
bands that can be studied in more detail in physiological studies, and more generally,
showcases a data driven approach to selecting frequency bands for analysis.
Chapter 5

Risk Stratification Using Health Insurance Records: Transferring Knowledge from Text

5.1 Introduction

Chapters 3 and 4 focused on a specific approach to feature engineering. In this chapter, we propose a method to transfer knowledge from text.

In many domains such as medicine, training data is in short supply. The need to use inclusion or exclusion criteria to select the population of interest, and the scarcity of many outcomes of interest further shrink the available data and compound the problem. The lack of data hinders learning of accurate risk models. A common technique in these scenarios is to leverage transfer learning from source data for related prediction tasks or populations [52, 102, 41]. Transfer learning allows knowledge such as the relative importance of each feature to be transferred from external sources to the task of interest. Although such approaches can improve the accuracy of risk models for the target task, obtaining enough useful source data is often difficult.

By contrast, free text corpuses such as Wikipedia are publicly available and may contain knowledge about the prediction task, such as predictive features that are
associated with the outcome of interest. Natural language processing of free text to learn these concepts is non-trivial and is an active area of research [24]. Recently, distributed representations of words such as word2vec have been shown to quantify similarities between words [62]. For example, “stroke” and “hypertension” have higher similarity compared to “stroke” and “fracture”. Because the first pair of words are both cardiovascular diseases but the second pair are less related, we reasoned that these similarity measures may also convey information about the relatedness of various concepts. We therefore hypothesized that these representations could be leveraged to quantify the relationship between the text description of each feature and the description of the outcome. These learned relationships could then be used to improve predictive models when training data are lacking.

In this work, we propose a novel method to transfer knowledge from text corpuses to improve the accuracy of predictive modeling using small datasets, as summarized in Figure 5-1. We use a type of shallow neural network model, word2vec, to learn from text. When trained on a large database of published biomedical articles, the word2vec model learns to maps biomedical terms to a numerical vector termed a
word vector. Next, we use the model to process the text descriptions of the features and outcome to estimate the relevance of each feature to predicting the outcome. Finally, we use these estimated relevances to inform the machine learning process, using a variant of a L1-regularized logistic regression, the adaptive lasso. Our method substantially improved the Area Under Curve (AUC) by 0.02 to 0.06 when there were few (25 to 50) positive examples. Moreover, our method selected 20 instead of 50 features on average, allowing experts to more easily interpret and use the model.

The remainder of the chapter is arranged as follows. We first describe related work in transfer learning and related supervised learning methods. Next, we describe the dataset and outcomes that we examine. Then, we describe our method and our experimental set up. Finally, we show results and discuss the potential relevance of our work to other domains.

## 5.1.1 Related Work

Our work generally falls under the rubric of transfer learning, which leverages data from a related prediction task, the source task, to improve prediction on the target task. Transfer learning has been productively applied to medical applications such as adapting surgical models to individual hospitals [52], enhancing hospital-specific predictions of infections [102], and improving surgical models using data from other surgeries [41]. A review of transfer learning can be found in [74].

Similar ideas in leveraging knowledge encoded in another database have been applied to clinical text classification, where the goal is to label patient records. For example, some authors used expert-coded ontologies such as the Unified Medical Language System (UMLS) to engineer and extract features from clinical text and used these features to classify the presence or absence of cardiovascular-related diseases [38] [105]. Others have also used ontology in the Open Directory Project [36] and Wikipedia [37] to generate features for text classification in non-clinical settings.

Our work diverges from prior work by transferring knowledge across different data types and across non-related tasks. We transferred knowledge from models trained on free text corpuses to predict each word’s context (i.e., the surrounding words),
to help identify relevant features for predicting clinical outcomes using structured medical data. We computed similarities between each feature’s text description and the outcome’s text description to estimate the relevance of each feature to predicting the outcome. We then used these relevance estimates to rescale the feature matrix, which is equivalent to controlling the relative regularization strength of each feature, as detailed in the next section.

This rescaling procedure is similar to the adaptive lasso [107] and its special case, the nonnegative garotte [22]. These were originally formulated for linear regression and led to a higher probability of selecting the truly relevant variables:

\[
\hat{w} = \arg \min_w \sum_j \frac{|w_j|}{r_j} + C \sum_i (y_i - w^T x_i)^2
\]

where \(i\) indexes patients, \(j\) indexes features, \(w\) is a vector of weights for the features, \(r\) is a vector of non-negative scaling factors for the weights, \(x_i\) is the feature vector for each patient, and \(y_i\) is the variable that we want to predict for each patient.

This equation can be reformulated such that the rescaling is performed on the original data matrix. First, define a new weight vector as the rescaled version of the original weight vector: \(\tilde{w}_j = \frac{w_j}{r_j}\). In the loss function (right summation above),

\[
w^T x_i = \sum_j w_j x_{ij} = \sum_j \tilde{w}_j r_j x_{ij} = \sum_j \tilde{w}_j \tilde{x}_{ij} = \tilde{w}^T \tilde{x}_i
\]

where \(\tilde{x}_{ij} = r_j x_{ij}\) is the original feature rescaled by \(r_j\). We can be solve for \(\hat{\tilde{w}}\) using the algorithms for solving the original, non-scaled expression [107]:

\[
\hat{\tilde{w}} = \arg \min_{\tilde{w}} \sum_j |\tilde{w}_j| + C \sum_i (y_i - \tilde{w}^T \tilde{x}_i)^2
\]
The learned \( \hat{\mathbf{w}} \) is the rescaled version of the original weight: \( \hat{w}_j = \hat{\mathbf{w}}_j r_j \). We apply this technique to logistic regression, replacing the loss function (right summation) to yield:

\[
\hat{\mathbf{w}} = \arg \min_{\mathbf{w}} \sum_j \frac{|w_j|}{r_j} + C \sum_i \log(1 + e^{-y_i \mathbf{w}^T \mathbf{x}_i})
\]

In the adaptive lasso, the adaptive scaling factors are usually obtained from the ordinary least squares estimate, which are weights trained on the same data without the regularization parameter (left summation). By contrast, we inferred these scaling factors from an expert text corpus. This technique makes use of auxiliary data and can thus be applied when the original training data are not sufficient to obtain reliable least squares estimates.

5.2 Data & Features

5.2.1 Data

Taiwan has had universal health insurance since 1995. Recently, the Government has started releasing subsets of the insurance records to Taiwanese Universities for research purposes. We have been fortunate to work on these datasets through collaborations with National Cheng Kung University in Tainan, Taiwan. For the research reported here, we used the Taiwan National Health Insurance Research Database, longitudinal cohort 3\textsuperscript{rd} release \cite{48}. The dataset contains all billing records for one million patients from 1996 to 2012. However, because the dataset used a Taiwan-specific diagnosis code (A-code) before 2000 instead of the International Classification of Diseases 9\textsuperscript{th} Revision, Clinical Modification (abbreviated ICD-9 henceforth) \cite{6}, we only utilized data after 2000.

It may seem odd to be using a dataset with one million patients to illustrate a technique designed for small datasets. However, we wanted to understand to what extent results achieved by applying our method to small number of examples could approximate results achieved using a conventional method on a very large number of
examples. In addition, we try to predict several rare outcomes that have only about one hundred examples, even within this large dataset.

We will detail the outcomes studied in the experiments section. For each outcome, we used five years of data (2002-2007) to extract features and the next five years (2007-2012) to define the presence of the outcome. We chose to use five-year periods based on the duration of available data and the availability of five-year risk calculators for related outcomes [59].

5.2.2 Feature Transformation

For our purposes, the raw billing data was a list of tuples \((date, billing code)\) for each patient. Because we summarized each patient’s data over a period of five years, the patient may have had multiple claims for any given billing code. Let the count of claims for patient \(i\), billing code \(j\) be \(z_{ij}\). We empirically found that the method of representing these counts \((z_{ij})\) in the features matrix \((x_{ij})\) significantly impacts the prediction performance. Using the raw counts \((x_{ij} = z_{ij})\) performs poorly relative to using a binary indicator function: \(x_{ij} = I(z_{ij} > 0)\). Transforming the counts using a log function was a further improvement: \(x_{ij} = 1 + \log(z_{ij}) \) if \(z_{ij} > 0\), and 0 otherwise. We normalized the log-transformed value to \([0, 1]\) based on the range of values observed in the training set. The same normalization procedure was used for the raw counts approach in our empirical studies.

We tested these different transforms because we hypothesized that the counts contained information about the severity of a condition. For example, a person with higher counts of stroke-related billing codes may have a more severe disease that requires more medical care. A \(\log(z)\) function increases quickly with small values of \(z\) but the gradient tapers off with higher values. However, the raw counts increases linearly with the counts (not identically because of normalization). The relative performance \((raw \ counts < binary \ indicator < log)\) indicates that most of the information content of the count of these billing codes are contained in first few counts.

To further explore this, we looked into other transformations that interpolate
between log-like shapes and sigmoid shapes. Specifically, we modified the Hill equation in biochemistry \cite{16}. The original equation was used in a completely different scenario, to characterize the binding of multiple copies of a molecule B to another molecule A. We defined the hill transform (Figure 5-2) as:

\[
\text{hill}\{x_{ij}\} = \frac{z_{ij}^p}{z_{ij}^p + m_j^p}
\]

where \(m_j\) is a parameter unique to each feature, that controls the “mid-point” of the function: \(\text{hill}\{x_{ij}\} = 0.5\) when \(z_{ij} = m_j\). We defined \(m_j\) to be the median across the non-zero counts of each feature. \(p\) is an exponent that controls the shape. \(p \leq 1\) creates a log-like function that increases quickly with small values of \(z_{ij}\) but tapers off as \(z_{ij}\) increases beyond \(m_j\); \(p > 1\) creates sigmoid shapes that increase slowly with small (or large) values of \(z_{ij}\), and quickly with values of \(z_{ij}\) close to \(m_j\).

We empirically found that most values of \(p\) were associated with better predictive performance relative to binary. Both the hill transform and the log transform taper the effect at high values of \(z_{ij}\), but hill also tapers the effect of the initial rise with \(p > 1\). However, this additional effect provided no value in our experiments.

Together, our data suggest that risk does not rise proportionately: the risk dif-
### Table 5.1: Examples of Billing Codes.

<table>
<thead>
<tr>
<th>Category (ICD-9)</th>
<th>Description of hierarchy level</th>
<th>Example code</th>
<th>Example description</th>
<th># of codes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diagnoses</td>
<td>Groups of 3-digit ICD-9</td>
<td>390-459</td>
<td>Diseases of the circulatory system</td>
<td>160</td>
</tr>
<tr>
<td></td>
<td>3-digit ICD-9</td>
<td>410</td>
<td>Acute myocardial infarction</td>
<td>1,018</td>
</tr>
<tr>
<td></td>
<td>4-digit ICD-9</td>
<td>410.0</td>
<td>(as above) of anterolateral wall</td>
<td>6,442</td>
</tr>
<tr>
<td></td>
<td>5-digit ICD-9</td>
<td>410.01</td>
<td>(as above) initial episode of care</td>
<td>10,093</td>
</tr>
<tr>
<td>Procedures</td>
<td>Groups of 2-digit ICD-9</td>
<td>35-39</td>
<td>Operations on the cardiovascular system</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>2-digit ICD-9</td>
<td>35</td>
<td>Operations on valves and septa of heart</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>3-digit ICD-9</td>
<td>35.2</td>
<td>Replacement of heart valve</td>
<td>890</td>
</tr>
<tr>
<td></td>
<td>4-digit ICD-9</td>
<td>35.21</td>
<td>Replacement of aortic valve w/ tissue graft</td>
<td>3,661</td>
</tr>
</tbody>
</table>

#### Medications (ATC)

<table>
<thead>
<tr>
<th>Anatomical main group</th>
<th>Chemical substance</th>
<th>Pharmacological subgroup</th>
<th>Therapeutic subgroup</th>
<th>Chemical subgroup</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>C03DA</td>
<td>C03D</td>
<td>C03</td>
<td>C03DA01</td>
</tr>
<tr>
<td>Cardiovascular system</td>
<td>Spironolactone</td>
<td>Potassium-sparing agents</td>
<td>Diuretics</td>
<td>Aldosterone antagonists</td>
</tr>
<tr>
<td>14</td>
<td>540</td>
<td>212</td>
<td>85</td>
<td>1,646</td>
</tr>
</tbody>
</table>

The difference between a count of 0 and 1 is greater than the difference between a count of 100 and 101. This is consistent with the fact that many risk scores use binary indicators as features with satisfactory predictive accuracy (Section 2.1.3). In addition, physicians also place greater emphasis on the presence or absence of a condition when summarizing a patient’s condition: language such as “patient has a history of hypertension, diabetes...” is commonly used.

Because the information retrieval field also uses counts (of words in a text document) as features, we also tested feature transformations such as variants of term-frequency inverse-document-frequency (TF-IDF) [11]. In fact, both the binary and log approaches are types of TF. Again, we empirically found that TF-IDF performed no better than log (results not shown), and thus used the log transformation.

### 5.2.3 Features and Hierarchies

Our features consisted of age, and billing codes that are either ICD-9 diagnosis and procedure codes [6], or medications. ICD-9 is the most widely used system of coding diagnosis and procedures in billing databases [27]. ICD-9 codes are arranged as a forest of trees, where nodes closer to the root represent more general concepts and nodes closer to the leaves represent more specific concepts. For example, the root

---

1Because we use a logistic regression model, the correct statistical term is odds; each unit increase of a feature increases the odds by an amount dictated by the weight of that feature.
node 410 indicates a diagnosis of acute myocardial infarction, the child node 410.0 states location of the infarction (Table 5.1). An illustration of these hierarchies can be found in Figure 2-8. We used these descriptions to estimate relatedness between each feature and the outcome.

Feature representations that exploit the hierarchical relationships of ICD-9 codes have been shown to improve predictive performance when using L2-regularized logistic regression [87]. We modified the propagated binary approach proposed by the authors to handle counts. In the original approach, all nodes are binary, and a node is 1 if either that node or any of its descendant are 1. In our modified “propagated counts” approach, the count of billing codes at each node was defined as the sum of the counts from itself and all of its descendants. The log transformation was applied after these counts were summed.

We expressed medications as a hierarchy by mapping the dataset’s medication codes to the World Health Organization Anatomical Therapeutic Chemical (ATC) Classification System [7]. Similar to the ICD-9 codes, ATC codes are arranged as a forest of trees where upper levels indicate categories of drugs and each node at the lowest level is a specific drug. Feature definitions and the potential number of features are listed in Table 5.1. We removed all features that were present in less than three patients in each training set, leaving approximately 10,000 features for each outcome.

In summary, each patient’s feature vector is of length 10,000. The first element of the vector is patient’s age in 2007. The other elements represent the counts of various billing codes in the five-year period (2002-2007) and may correspond to diagnoses, procedures, or medications.

5.3 Methods

5.3.1 Computing Estimated Feature-Relevance

Figure 5-1 summarizes our approach. We start with a word2vec model trained on a corpus of medical articles, and use this model to compute the relevance of each
The rescaled features in the final step are then used as input to a L1-regularized logistic regression model.

Word2vec Model

Word2vec uses a corpus of text to produce a low dimensional dense representation $\in \mathbb{R}^d$ for each word, i.e., word vectors [62]. For example, as illustrated in Figure 5-4, a word is most easily represented in a one-hot encoding, which is a long sparse binary vector that has length equal to the size of the vocabulary. Word2vec and other similar techniques map this sparse vector to a lower dimensional dense vector.
We used word2vec instead of other vector representations because of its computational efficiency: high quality vectors can be learned from a dataset of 1.6 billion words in a day. We used the skip-gram variant of word2vec, which is trained to predict the context for each word, or the surrounding words in the corpus (Section 2.2.3). Thus words used in similar contexts are represented by similar vectors, and these vectors could be used to quantify similarity.

We used models pre-trained on 22,723,471 PubMed article titles and abstracts and 672,589 PubMed Central Open Access full text articles [79]. These models used a skip-gram model with \( d = 200 \) dimensional vectors and a window size of 5. The window size is a parameter that defines the distance of the context; this model learns to predict words that are within a window of 5 words. The learned vectors were useful in identifying related biochemical terms, such as different types of amino acids, and different types of biochemical modifications for DNA (deoxyribonucleic acid). The vectors could also be used to help identify diseases mentioned in biomedical text articles despite variations in exact words used. We used the learned word2vec vectors to estimate the relevance of each features to predicting the outcome. The following sections detail this process, and an example of this procedure applied to a single feature can be found in Figure 5-3.

**Construction of Hierarchical Feature Descriptions**

Some ICD-9 descriptions are not informative without knowledge of its ancestors. For example, the description for ICD-9 014 is “Tuberculosis of intestines, peritoneum, and mesenteric glands,” but its child node ICD-9 014.8 has the description “Others.” To ensure that each feature was assigned an informative description, we concatenated each feature description with the descriptions of its ancestors in the hierarchy. For example, diagnosis code 401 (Essential Hypertension) had as its ancestor the node representing the group of codes 401-405 (Hypertensive Disease), which in turn had as its ancestor 390-459 (Diseases of the Circulatory System). Thus the hierarchical feature description for code 401 was the concatenated description for these three nodes (Figure 5-3). We termed these concatenated descriptions “hierarchical feature
descriptions.”

**Computing Similarity for Each Word in Feature**

Next, our goal was to estimate the relevance of each hierarchical feature description to the outcome. First, we measured the similarity between each word in a hierarchical feature description and the outcome. Next, we summarized these similarity values to obtain a single scalar.

To obtain the similarities values, we first removed extraneous or “stop words” based on a standard English list [78], augmented with the words “system,” “disease,” “disorder,” and “condition.” These augmented words were chosen based on frequently occurring words in our feature descriptions. For example, the hierarchical description for code 401 described in the previous section was filtered to remove the words “diseases,” “of,” “the,” “system,” and “disease”. The remaining words were “circulatory,” “hypertensive,” “essential,” and “hypertension.” We applied the Porter2 stemmer [78] to ensure that each word in the descriptions was matched to the closest word available in the word2vec model despite different word endings such as “es” and “ed.”

Next, we computed the similarity between each word in the filtered hierarchical feature description and the outcome description. For simplicity, each outcome was summarized as a single keyword, e.g., “diabetes” (Table 5.2).

For the filtered description for code 401 for example, we computed the similarity between “circulatory” and “diabetes” by extracting the two word vectors, vec(“circulatory”) and vec(“diabetes”) from the word2vec model. As standard for word2vec models, we used the cosine similarity to quantify the similarity between the two vectors. The cosine similarity is the dot product of two normalized word vectors, and lies in the range \([-1, 1]\]. We rescaled each similarity value to the range \([0, 1]\) to obtain non-negative scaling factors for our next step.\(^2\) We then repeated this procedure for each of the remaining words. We obtain four similarity values in this example: 0.56, 0.70, 0.54, and 0.83 (Figure 5-3).

\(^2\)We justify this method of rescaling (mapping -1 to 0 and 0 to 0.5) by considering a hypothetical pair of words, \(w_1\) and \(w_2\), whose normalized word vectors are negations of each other: \(\text{vec}(w_1) = -\text{vec}(w_2)\). Thus their cosine similarity is \(-1\). When used in the word2vec model, these vectors
Averaging Word Similarities

Recall that each hierarchical feature description of each feature contained multiple words, which were filtered and then converted to similarity values, one for each filtered word. The presence of a single word may be sufficient to indicate that the feature was relevant. For example, because diabetes is known to be strongly associated with hypertension, a feature that contains the word "hypertension" in its description is likely to be relevant. Mathematically, this would be equivalent to taking the maximum of the values in the similarity vector. However, the maximum function ignores all other words in the feature description and prevents relative ranking between features that contain the same word. For example, "screening for hypertension" and the actual diagnosis code for hypertension will receive the same relevance estimate despite the fact that the former description does not actually indicate that the patient has hypertension. Furthermore, the maximum function is sensitive to outliers, and thus features may be assigned erroneously high relevances.

On the other hand, taking the arithmetic mean of the similarity vector would effectively erase the effect of any words that truly indicate high relevance. Thus, we selected the power mean, which is biased towards the maximum, but takes all the words into account. The power mean is defined as $(\frac{1}{m} \sum_i s_i^p)^{\frac{1}{p}}$, where $s_i$ is the $i^{th}$ similarity value out of $m$ filtered words in the hierarchical feature description. $p$ is a tunable exponent that can interpolate the function between the maximum ($p = +\infty$) and the arithmetic mean ($p = 1$). Thus if the hierarchical feature description contained a few words with high similarity to the outcome, the averaged similarity scalar was high as well.

have opposite predictions because of the opposite signs; words that are likely to appear near $w_1$ are less likely to appear near $w_2$ and vice versa. Thus if the outcome description is $w_1$ and the feature description is $w_2$, this feature should be irrelevant in predicting the outcome, and thus be mapped to a relevance of 0.

Consistent with this intuition, we find that the lowest cosine similarities (before rescaling) between pairs of words are approximately -0.3. In other words, all the words in our text corpus are somewhat related to each other. This can be rationalized by considering the fact most words occur within some distance of common words such as "the" and "a", and thus, all words share some context.

The proof of this is as follows: let the maximum $s_i$ be $s_{max}$ (so $s_i \leq s_{max}$). Then,
Figure 5-5: Histograms of feature relevances computed using various means: arithmetic mean, power means with exponents 10 and 100, and maximum values.

We chose the exponent $p$ by plotting the distributions of feature relevances when computed using $p = 1, 10, 100, +\infty$ (Figure 5-5). The $p = 100$ curve tracks the maximum closely, and may be overly sensitive to the maximum value in each similarity vector, whereas the $p = 10$ curve “pushes” an additional 10% of the features towards higher relevances relative to the arithmetic mean. Because this reflected our expectation that relatively few features were closely related to the outcome, we selected the power mean with an exponent of 10.

Taking the power mean of the similarity vector outputs a single scalar that quantified the similarity of that feature with the outcome. For example, the power mean of the similarity vector for code 401 was 0.74 (Figure 5-3).

Rescaling Features based on Similarity

Finally, we multiplied each feature value by its estimated feature-relevance. This changed the numerical range of the feature from $[0, 1]$ to $[0, relevance]$. This is equivalent to:

$$\lim_{p \to \infty} \left( \frac{1}{m} \sum_i s_i^p \right)^{\frac{1}{p}} = \lim_{p \to \infty} \left[ \frac{1}{m} \sum_i \left( \frac{s_i}{s_{max}} \right)^p s_{max}^p \right]^{\frac{1}{p}}$$

$$= \lim_{p \to \infty} \left[ s_{max}^p \frac{1}{m} \sum_i \left( \frac{s_i}{s_{max}} \right)^p \right]^{\frac{1}{p}} = \lim_{p \to \infty} s_{max} \left[ \frac{1}{m} \sum_i \left( \frac{s_i}{s_{max}} \right)^p \right]^{\frac{1}{p}}$$

For $s_i \neq s_{max}$, $\lim_{p \to \infty} \left( \frac{s_i}{s_{max}} \right)^p = 0$. Thus the summation reduces to counting the number of $s_i$ that are equal to $s_{max}$, yielding:

$$\lim_{p \to \infty} s_{max} \left[ \frac{1}{m} \sum_i I(s_i = s_{max}) \right]^{\frac{1}{p}}.$$  
Because the term in the square braces $\leq 1$, this reduces to $s_{max}$.  

100
Table 5.2: Definitions of outcomes and keyword used in estimating the relevance of each feature. CVA= cerebrovascular accident (stroke); CHF= congestive heart failure; AMI= acute myocardial infarction; DM= diabetes mellitus; HCh= hypercholesterolemia.

<table>
<thead>
<tr>
<th>Outcome (keyword)</th>
<th>Billing codes (definition)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CVA (cerebrovascular)</td>
<td>ICD-9 433 (Oclusion and stenosis of precerebral arteries)</td>
</tr>
<tr>
<td></td>
<td>ICD-9 434 (Oclusion of cerebral arteries)</td>
</tr>
<tr>
<td></td>
<td>ICD-9 436 (Acute, but ill-defined, cerebrovascular disease)</td>
</tr>
<tr>
<td>CHF (CHF)</td>
<td>ICD-9 428 (Heart failure)</td>
</tr>
<tr>
<td>AMI (coronary)</td>
<td>ICD-9 410-414 (Ischemic Heart Disease)</td>
</tr>
<tr>
<td>DM (diabetes)</td>
<td>ICD-9 250 (Diabetes mellitus)</td>
</tr>
<tr>
<td></td>
<td>ATC A10 (Drugs used in diabetes)</td>
</tr>
<tr>
<td>HCh (hypercholesterolemia)</td>
<td>ICD-9 272.0 (Pure hypercholesterolemia)</td>
</tr>
<tr>
<td></td>
<td>ICD-9 272.1 (Pure hyperglyceridemia)</td>
</tr>
<tr>
<td></td>
<td>ICD-9 272.2 (Mixed hyperlipidemia)</td>
</tr>
<tr>
<td></td>
<td>ICD-9 272.3 (Hyperchylymicronemia)</td>
</tr>
<tr>
<td></td>
<td>ICD-9 272.4 (Other and unspecified hyperlipidemia)</td>
</tr>
<tr>
<td></td>
<td>ATC C10 (Lipid modifying agents)</td>
</tr>
</tbody>
</table>

alient to using the adaptive lasso \[107\] with an adaptive weight of \(1/relevance\). Intuitively, for features with low estimated relevance, this effectively required a larger feature weight to contribute to the prediction output.

## 5.4 Experiment Setup

### 5.4.1 Outcomes

We applied our method to predicting the onset of various cardiovascular diseases. We focused on five adverse outcomes: cerebrovascular accident (CVA, stroke), congestive heart failure (CHF), acute myocardial infarction (AMI, heart attack), diabetes mellitus (DM, the more common form of diabetes), hypercholesterolemia (HCh, high blood cholesterol).

The definitions that we used for each outcome are shown in Table 5.2. For each outcome, we used five years of data (2002-2007) to predict the presence of the outcome.
in the next five years (2007-2012). Based on advise from our clinical collaborators, we defined each outcome as three occurrences of the outcome’s ICD-9 code or the code for the medication used to treat the disease. To help ensure that we were predicting the onset of each outcome, we excluded patients (from both training and test sets) that had at least one occurrence of the respective billing codes in the first five years (Table 5.2, Figure 5-6).

We built separate models for two age groups: (1) ages 20 to 39, and (2) ages 40 to 59 (Table 5.3) and for males and females. We separated the age groups because the incidence of the outcomes increased with age. The age 40 was chosen to split our age groups because 40 is one of the age cutoffs above which experts recommend screening for diabetes [89]. Patients below 20 and above age 60 were excluded because of very low or high rates of these outcomes, respectively. In addition, we built separate models for males and females because differences in male and female physiology lead to many medications and diagnoses being strongly correlated with gender. For example, the incidence of all the diseases studied in this chapter are higher in males than in females (Table 5.3). In summary, we built models for five diseases, two genders, and two age groups for a total of 20 prediction tasks.

5.4.2 Experimental Protocol

For each prediction task, we split the population of patients into training and test sets in a 2:1 ratio, stratified by the outcome to ensure equal ratios of outcomes in the training and test sets. In the training set, we learned the weights for a L1-regularized logistic regression model using the implementation in liblinear [34]. The cost parameter was optimized by two repeats of five-fold cross validation on the
Table 5.3: Prediction tasks and number of patients. Acronyms as per Table 5.2.

<table>
<thead>
<tr>
<th>Outcome -gender</th>
<th>Age group 1: 20 – 39</th>
<th>Age group 2: 40 – 59</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>n (%)</td>
</tr>
<tr>
<td>CV A-F</td>
<td>171836</td>
<td>100 (0.1%)</td>
</tr>
<tr>
<td>CV A-M</td>
<td>156649</td>
<td>224 (0.1%)</td>
</tr>
<tr>
<td>CHF-F</td>
<td>171749</td>
<td>137 (0.1%)</td>
</tr>
<tr>
<td>CHF-M</td>
<td>156704</td>
<td>217 (0.1%)</td>
</tr>
<tr>
<td>AMI-F</td>
<td>170247</td>
<td>552 (0.3%)</td>
</tr>
<tr>
<td>AMI-M</td>
<td>154921</td>
<td>1109 (0.7%)</td>
</tr>
<tr>
<td>DM-F</td>
<td>168194</td>
<td>1609 (1.0%)</td>
</tr>
<tr>
<td>DM-M</td>
<td>153834</td>
<td>2273 (1.5%)</td>
</tr>
<tr>
<td>HCh-F</td>
<td>166938</td>
<td>2386 (1.4%)</td>
</tr>
<tr>
<td>HCh-M</td>
<td>149061</td>
<td>5008 (3.4%)</td>
</tr>
</tbody>
</table>

training set. Because only 0.1% to 12% of the population experienced the outcome in each prediction task, we set the second parameter (the asymmetric cost parameter) to the class imbalance ratio. We repeated the training/test split 10 times, and report the area under receiver operating characteristic curve (AUC) \[44\] averaged over the 10 splits.

To assess the statistical significance of differences in the AUC for each task, we used the nonparametric sign test, which uses the binomial distribution to quantify the probability that at least \( n \) values in \( m \) matched pairs are greater \[82\]. This requires no assumptions about the distribution of the AUC over test sets, or the magnitude of improvement. When we compared two methods using this test, a significant \( p \)-value (\( p<0.05 \)) was obtained when one method was better in at least 9 out of the 10 test sets.

5.5 Experiments and Results

In this section, we will describe our experiments and results. We first describe the computed feature similarities and the types of features at the top and bottom of the ranked list. Next, we show the results of training models on the full dataset. Finally, we show results of training models on subsampled data to evaluate our method when less training data are available.
5.5.1 Ranking of Features

We first verified that the word2vec similarity and our power law averaging provided a sensible relative ranking of features with respect to each outcome. For the AMI (heart attack) outcome for example, highly ranked features descriptions were related to the heart, with a similarity ranging from 0.65 to 0.85. The least similar feature descriptions, such as skin disinfectants and cancer drugs, had similarities close to 0.5.

5.5.2 Full Dataset

Because there was a wide range in number of outcomes in our tasks, we reasoned that evaluating our method on the full dataset would provide information about the regime of data size for which our method might be useful. Thus, we first trained models for each of the 20 tasks and assessed their performance. Our proposed method meaningfully outperformed the standard approach in one task (AUC of 0.718 compared to 0.681 for CVA in age group 1 of females). We had the fewest number of positive examples in the training data (67) for this task. Any statistically significant differences in the remaining tasks were small, ranging from −0.009 to 0.009. This indicated that our method had minimal effect on performance (either positive or negative) when data were plentiful.

In 13 out of 16 tasks where the number of selected features were statistically different, our method selected fewer features. On average, our method selected 83 features, compared to 117 for the standard approach.

Because the most significant difference and several non-statistically significant differences occurred in predictive tasks with less than 100 outcomes in the training set, we hypothesized that our method would be most useful in situations with few positive examples (<100).

5.5.3 Downsampled Dataset with 50 Positive Examples

To test our hypothesis, we used the same splits as in the previous section, and down-sampled each training split such that only 50 positive examples and a proportionate
number of negative example remained, thereby preserving the class imbalance ratio. We assessed the effect of training on this smaller training set, and computed the AUC averaged across the same (larger) test sets and evaluated significance in the same way.

Our proposed method yielded improvements relative to the standard approach in 10 out of 20 tasks, with differences ranging from 0.020 to 0.065 (Figure 5-7). The remaining 10 tasks respectively did not have statistically significant differences. Among the tasks with significant differences, our method had an average AUC of 0.697, compared to 0.656 for the standard approach. In addition, where the number of selected features using 50 positive examples using the standard approach compared with our proposed rescaling approach. Outcome and gender are labeled on the x-axis, each spanning two predictive tasks: age groups 1 and 2. Triangles indicates a significantly higher (pointed up) or lower (pointed down) AUC or number of selected features for our proposed method. Error bars indicate standard error.
features were statistically different, our method selected 20 features, compared to 50 for the standard approach.

5.5.4 **Downsampled Dataset with 25 Positive Examples**

After we further downsampled the training data to have 25 positive examples, similar observation were made: our method yield improvements in 6 tasks, with differences ranging from 0.022 to 0.086. The remaining 14 tasks did not have statistically significant differences. Among the tasks with significant differences, our method had an average AUC of 0.658 compared to 0.606 for the standard approach. Similarly, where the number of selected features were statistically different, our method selected 52 features, compared to 80 for the standard approach.

A summary of results from all three experiments are in Figure 5-8, showing the overall trend of improved performance and fewer selected features when comparing our method to the standard approach.

5.6 **Discussion**

In this chapter we demonstrated a method to leverage auxiliary, publicly available free text data to improve classification of adverse patient outcomes using a separate,
structured dataset. We showed that this method can improve prediction performance, particularly in cases of small data. To our knowledge this is the first work that transfers knowledge in this manner.

Furthermore, our method consistently selects fewer features (20 vs 50) from an original feature dimensionality of 10,000. This allows domain experts to more easily interpret the model, which is a key requirement for medical applications to see real world use.

To better understand the reason for our improved performance and number of selected features, we plotted the number of selected features against the estimated feature-relevance. The plots for predicting stroke in patients aged 20-39 are shown in Figure 5-9 females on the left and males on the right. These plots were representative of other tasks where the performance was statistically different. In females, where the method improved performance and substantially decreased the number of selected features, there is a marked decreased in selected features with estimated relevance below 0.7. However, there is little change in the number of selected features with higher estimated relevance. In other words, our approach preferentially removed features that were estimated to be have low relevance. By removing these “noise” features, more accurate weights could be assigned to the remaining features and thus improve performance.

Diagnosis code 373.11 (Hordeolum externum) is an example of a feature that was selected by the original approach but not our method. This condition, also termed
a stye or sty, is a small bump on the surface of the eyelid caused by a clogged oil gland. This is not expected to be strongly related to stroke, and may have been selected because of noise in the data. Our approach assigns this feature an estimated relevance of 0.57, which decreased the probability of it being selected.

In males however, where neither the performance nor the number of selected features were meaningfully different, there was little difference in the estimated relevance of selected features. Among the 20 prediction tasks, the phenomenon of preferentially selecting fewer less relevance features occurred in almost all (9 out of 10) cases where performance improved.

The improvements in some prediction tasks but not others may be explained by our method of estimating feature-relevance. We estimated relevances by outcome and did not take into account the specific patient population (e.g., age group and gender). Therefore the estimates may be better suited to some patient populations. This is qualitatively supported by some trends: our method is an improvement in predicting CVA only in females; for predicting CHF and DM only in the older age group. However our method was an improvement for HCh only in the younger age group in males and an improvement in all populations except female in older age group for AMI. A more sophisticated method of estimating feature relevance that takes into account the patient population might improve the results.

An advantage of our method is its small number of parameters; the power mean exponent is the only tunable parameter, and this exponent can, in principle, be selected by cross validation. Because our method is not specific to medicine, it can also be applied to other domains where feature descriptions are available (as opposed to numerically labeled features). In the absence of a corresponding domain specific text corpus, a publicly available general corpus such as Wikipedia may suffice to estimate the relevance of each feature to the outcome.
5.7 Limitations and Future Work

Our work has several limitations. First, the data are comprised of billing codes, which may be unreliable for the purposes of defining the presence or absence of a disease. We have defined each outcome to be three or more occurrences of the billing code, however, this creates a gray zone. Patients who have had one or two occurrences are labeled as negative in our prediction task. An ideal solution would involve measurements of the clinical variable of interest (e.g., blood sugar or A1C for diabetes). Unfortunately, these measurements were not available in our dataset.

Next, our method used the word2vec model, which is trained to predict the context in which each word is used. We then used this model to measure the similarity between word vectors and averaged these similarity values to obtain an estimated relevance for each feature. We could probably improve this procedure using the correlation or co-occurrences of various billing codes to improve these estimates. For example, a recurrent neural network (RNN) \cite{47} could read the word vectors for the descriptions of features $i$ and $j$, and predict the correlation between the two features in the dataset. This RNN could potentially generalize to outcomes and feature combinations that it has not seen before, and thus be used for rare adverse outcomes in our dataset. This approach has the advantage of being able to handle syntax such as negation in each feature description, and handle outcome descriptions with multiple words. However, this approach requires additional data with which to train and refine the estimated relevance values, whereas our current approach does not.

One may wonder if the co-occurrence of words can be directly used to compute similarity: if two words frequently occur in the same document or context, then these words are related. We believe that this approach would be problematic because of indirect effects. For example, diabetes and hypertension are related because of shared risk factors. Thus we might expect “diabetes” and “hypertension” to co-occur more frequently than might be expected by random chance. However, we would also like for equivalent concepts (such as drugs used to treat these conditions) to be considered related. Because biomedical text that mention “diabetes” (or “hypertension”) are
unlikely to contain a comprehensive list of all the related drugs, a simple co-occurrence approach might miss these “indirect” relationships. Though it has yet to be formally proved, the empirical data suggest that word vectors tend to map these words to similar vectors [53]: if the respective drugs are mentioned together with “diabetes” (or “hypertension”) in other contexts, these indirect relationships could be learned.

Finally, we have used our feature relevances and the adaptive lasso to aid feature selection. However, this general framework of estimating the relevance of each feature from external data could also be used with L2 regularization, where it will preferentially shrink less relevant features’ weights towards zero. This may result in more accurate weight estimates (and thus improve performance), but without a feature selection effect.

5.8 Conclusion

We have demonstrated an approach to transfer knowledge from the medical literature to improve the accuracy of predictive models trained on small datasets. Our approach uses the text descriptions of the features to estimate the relevance of each feature towards predicting the outcome. We then use these relevance estimates to help select the most relevant features when training the model. We show that this improves model performance. This approach can also be applied to other clinical predictive tasks where text descriptions of the features and outcome are available.
Chapter 6

Summary and Conclusion

We end with a summary of our methods to leverage medical knowledge in predictive modeling, and a broader discussion of implications and future work.

6.1 Feature Engineering for the ECG

Chapters 3 and 4 used beat-frequency, which was designed to measure frequency with respect to cardiac cycles, to create better features. The first step involves extracting a time series of interest from the ECG, where there is one value in the new time series for each heartbeat in the ECG. Next, we estimated the power spectrum of this new times series. We then learned weights for these features either by testing all possible combinations of frequency bands (Chapter 3) or by using machine learning (Chapter 4).

In Chapter 3, we applied our approach to the morphologic distance time series, which quantifies the beat-to-beat difference in ECG morphology. When evaluated on 1,082 patients after an non-ST-elevation acute coronary syndrome (NSTEACS), our method was particularly accurate in patients that other clinical measures indicate as “low-risk”. The one-year hazard ratios in this low risk cohort (as determined by TIMI risk score, ejection fraction, and B-type natriuretic peptide) was 14.9, compared to 5.9 for the original risk metric in time-frequency. Despite the large difference, it was difficult to assess statistical differences because of the small numbers of adverse
events in this population. Moreover, our approach could group the beat-frequency spectra into bands with high intra-band correlation and low inter-band correlation for subsequent analysis. This delineation was less clear in time-frequency.

In Chapter 4, we applied our approach to the instantaneous heart rate time series, which is measured in terms of the NN time interval in milliseconds. When evaluated on 2,302 patients post NSTEACS, our method yielded significant differences in AUC when using pre-defined frequency bands from the literature: 0.73 compared to 0.70 for time-frequency. Surprisingly, performance for the time-frequency metric did not improve with use of machine learning, whereas the beat-frequency AUC improved to 0.75. Our data indicate that this is because the location of time-frequency bands (in Hz) shift with heart rate changes, making it difficult to learn time-frequency bands that are prognostic in different patients (with different heart rates). By contrast, beat-frequency bands are less sensitive to heart rate changes. Our approach identified several beat-frequency bands for future studies.

### 6.1.1 Implications and Future Work

Beat-frequency is designed to extract frequency-domain information from the ECG. We have applied it to two time series derived from the ECG, one that quantifies beat-to-beat differences in ECG morphology, and another that quantifies instantaneous heart rate. Beat-frequency can also be directly applied to other time series derived from the ECG. For example, T-wave alternans is obtained from the time series of ST-segment and T-wave amplitudes for each beat, and measuring variability at 0.5 cycles/beat. This could be extended to amplitude across the entire beat in each part of the ECG (not just the ST/T), and across the whole frequency spectrum. A machine learning approach similar to those we have described could use these data as features, to find ECG morphological changes in the frequency domain that are correlated with adverse events. Note that although this also quantifies changes in morphology, it is distinct from MVB and MV. MVB and MV are computed from the beat-to-beat morphological difference time series; each point corresponds to the difference in morphology between that beat and the next.
Similarly, beat-frequency can be applied to other cardiac signals, such as blood pressure waveforms, heart sound recordings from electronic stethoscopes, and cardiac imaging modalities such as real time cardiac echocardiogram (ultrasound) and Magnetic Resonance Imaging (MRI). A general framework would encompass the following steps. First, segment the signal (where images can be considered two-dimensional signals) so that every heartbeat is associated with one or more values. In the imaging scenario, this value could quantify the ejection fraction or stroke volume (volume of blood pumped) of each heartbeat. Next, compute the beat-frequency spectra of this derivative time series, and search for characteristic bands, or adopt a data-driven approach using patient outcomes as labels.

We are also intrigued in applications of this idea beyond the cardiac sphere. Other types of frequency domain could potentially be defined for other quasi-periodic signals, such as respiratory waveforms, human gait, and speech signals.

6.2 Feature Selection Using Knowledge from Text

Chapter 5 tackled the problems of prediction performance and selecting features, in the regime of small data. We leveraged knowledge from publicly available medical articles to estimate the relevance of each feature to the outcome. We then used these relevances to improve feature selection, and thus improve prediction performance. We applied our method to predicting the onset of various cardiovascular disease in different patient populations, resulting in 20 prediction tasks.

Our approach was particularly effective when there were < 100 positive examples in the training data. For example, when applied to training data that had 50 positive examples, our approach improved prediction performance in 10 out of the 20 prediction tasks, and had similar performance to the standard approach in the remaining tasks. Among tasks with a statistically different performance, our approach had an average AUC of 0.70, compared to 0.66 for the standard approach. In addition, our approach selected 20 instead of 50 features on average, allowing easier interpretation of the model by physicians and potentially easing the way for such models to see
real world use. Our data also indicate that the method is preferentially removing features that are estimated to have low relevance, and thus methods that improve this estimation procedure might further improve our results.

6.2.1 Implications and Future Work

To our knowledge, our method is the first demonstration of automatically transferring knowledge from text to estimate the relationship of the outcome to features extracted from structured data. This procedure is analogous to asking a domain expert (a physician in this case) about the relevance of each feature to the outcome. However because our method is automated, it can scale to arbitrarily large numbers of features (e.g., $10^5$), which is not possible with manual annotation.

This framework is flexible and can be applied to non-medical applications by replacing the biomedical text with either an application-specific text corpus, or a more general corpus such as Wikipedia. In addition, the process of learning knowledge from text can be improved. For example, ontologies such as Unified Medical Language System (UMLS) contain expert-coded knowledge that can guide the learning process. Imagine that a model has “read” several text articles that discuss both hypertension and diabetes. The ontology may inform the model that these two conditions are co-morbidities, i.e., patients with one condition are likely to have the other. Based on these data, the model could learn the syntax of natural language used to describe such relationships, and it may then learn other co-morbidities from other text articles.

Finally, an extension to estimating feature relevance would be estimating relative feature weight. These weight estimates could then be used as priors either in a Bayesian setting or by shrinking the weights towards this estimated weight. This is important because the current method that affects relative feature regularization does not allow the method to learn about features that are not present in the training set. By contrast, having weight estimates would allow estimation of the value of features that may not have been seen as predictive in the training set (because of noise in small datasets).
Bibliography


123


