Addressing Quality of Patient Care by Measuring Care Variation and Predicting Thirty-Day Readmissions

---

A Dissertation

Presented to

the faculty of the School of Engineering and Applied Science

University of Virginia

---

in partial fulfillment

of the requirements for the degree

Doctor of Philosophy

by

Michael A. Vedomske

May

2014

APPROVAL SHEET

The dissertation

is submitted in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Michael Vedomske

AUTHOR

The dissertation has been read and approved by the examining committee:

Donald Brown

Advisor

Gerard Learmonth

James Harrison Jr.

Matthew Gerber

Marc Breton

Accepted for the School of Engineering and Applied Science:

Dean, School of Engineering and Applied Science

May

2014

ADDRESSING QUALITY OF PATIENT CARE BY MEASURING CARE VARIATION AND PREDICTING THIRTY-DAY READMISSIONS

A PH.D. DISSERTATION

Michael Vedomske

Systems & Information Engineering

University of Virginia

Committee

Prof. Gerard Learmonth, Ph.D.– SIE—*Chair*
Prof. Donald Brown, Ph.D.– SIE—*Advisor*
Dr. James Harrison Jr., M.D., Ph.D.– PHS —*Outside Member*
Prof. Matthew Gerber, Ph.D.– SIE
Prof. Marc Breton, Ph.D.– PNBS

May 2014

Abstract

Despite industrialized nations' increases in medical investment, care organizations have been unable to consistently deliver quality healthcare to all patients. Healthcare decision makers must find ways to improve delivery during a patient's stay at a hospital while also assuring they are discharged at the proper time. Oftentimes, decision makers rely on care variation studies to understand trends in the consistency of patient care across many patient cohorts. Decision makers also measure the quality of transition from discharge to home through unplanned thirty-day readmissions. Unplanned thirty-day readmissions has become the legislated metric for quality of care in the United States. This dissertation provides metrics to measure and assess care organizations' variation in care both between and within patient cohorts. This dissertation also provides predictive models and modeling recommendations across a broad set of data, variable, and model choices to improve unplanned thirty-day readmission prediction. These two combined provide a set of care quality tools for addressing consistency in care during a visit and data driven decision tools for deciding when that visit should end.

This research provides six multidisciplinary contributions. First, this dissertation provides a data framework for simplifying and utilizing a complicated, high-dimensional data structure consisting of many categorical variables. Second, this dissertation demonstrates a statistically viable method for measuring variation between two columns of the data framework. Third, this research demonstrates metrics for measuring variation within columns of the framework and provides validation for their utility. Fourth, this dissertation assesses the performance and parameterization of three algorithms on high-dimensional data sets. Fifth, for high-dimensional data, this research assesses the utility of methods commonly used to address class imbalance. Finally, this research provides evidence that careful selection of variable representation when deriving new variables for predictive models has compounding effects on model performance.

This dissertation also provides six healthcare contributions. First, this research demonstrates the utility of our between cohort variation method across both principal and all procedures for 2,383 comorbidities as well as three additional cases. Second, this research provides evidence for increased chance of variation due to lack of diagnostic specificity. Third, this research derives more than a dozen new variable representations, including within cohort variation metrics, and presents their predictive performance of unplanned thirty-day readmissions. Fourth, this research has demonstrated a method for ranking procedures for analysts to explore in order to reduce thirty day readmissions and care variation. Fifth, this dissertation has developed the best performing model of thirty-day readmissions to date. Finally, this dissertation provides multiple recommendations for shifting the thirty-day readmissions modeling paradigm to markedly improve predictive performance.

Acknowledgements

I would like to acknowledge all who have played a role in my academic development over the last six years. Special thanks belong to Prof. Don Brown for taking me on even when he had so many students and for helping me find a problem space and data set with which to make a difference. Our conversations and his pointed guidance were instrumental in my research progress. Great thanks also belong to Dr. Jim Harrison who provided much food for thought, mentorship, medical expertise, and novel ideas. He always had an open door and listening ear. Thank you to Prof. Matt Gerber for his excellent questions and intellectual approach during the development of this research and keen eye while writing. Thanks as well to Profs. Learmonth and Breton for providing feedback and perspective as I've made adjustments throughout the research phase.

Others have been instrumental in my research progress. Ken Scully has been a tremendous help in obtaining, understanding, cleaning, and interpreting the data from the CDR. Without his help and ideas, this research may never have been completed. Thank you to Prof. Laura Barnes who provided me with perspective on the field of medical informatics.

Many students and others have provided encouragement, practical suggestions, and otherwise helped me along. Of note is Samuel Huddleston who was my sounding board on many occasions. His focus and well-placed advice saved me much trouble and helped shape my overall research trajectory. Ed Teague also provided much needed encouragement and friendship during my research. There are too many others to mention but thanks go out to them.

I would also like to thank Jennifer Mauller, Jayne Weber, Terri Corcoran, and Deb Hirst for their willing help throughout the years. They have helped in numerous ways and always with a smile.

This research would not have been possible without funding from the National Science Foundation through its Graduate Research Fellowship Program as well as generous funding multiple times from the Department of Systems and Information Engineering and the Graduate School of Engineering and Applied Sciences.

Finally, I thank the love of my life, and the greatest proofreader I've met, my dearest love and wife, Allison. She has been a constant support and strength from our undergraduate education all the way through my PhD research. I truly appreciate her sacrifice and support. My children have also been a strength throughout the last six years. They have provided me with excited shouts, hugs, and love as I've come home each day. These were especially appreciated after a long day of struggle in my research. This research has been the result of much pondering, hard work, and many answered prayers by my family and myself.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

> As medical science and technology has advanced at a rapid pace, the health care delivery system has floundered in its ability to provide consistently high quality care to all. [2]

According to a national review by the Institute of Medicine in 2006, despite improvements in medical understanding and in the technology used to directly treat patients, the health care systems in the United States have not been able to consistently provide care at the same level across its people. The World Health Organization points out that this indictment does not fall only on the US, but on many industrialized nations, where spending increases have not led to proportional increases in quality of health care [3].

The World Health Organization stated that "...[t]aking a systems perspective, and orienting systems to the delivery and improvement of quality, are fundamental

to progress" on this front [3]. This dissertation is a response to this call.

In a report to the US Congress in response the the Institute of Medicine (IOM) report, the Centers for Medicare and Medicaid Services (CMS) laid out a Quality Measurement Roadmap for improving quality of care [4]. Included in these plans were six quality measure goals:

1. **Safety** where care doesn't harm patients.
2. **Effectiveness** where care is evidence-based.
3. **Smooth Transitions of Care** where care is well-coordinated across different providers and settings.
4. **Transparency** where information is used by patients and providers to guide decision-making and quality improvement efforts, respectively.
5. **Efficiency** where resources are used to maximize quality and minimize waste.
6. **Eliminating Disparities** where quality care is reliably received regardless of geography, race, income, language, or diagnosis. [4]

This dissertation addresses concerns in goals three, Smooth Transitions in Care, and six, Eliminating Disparities. We address goal three by exploring issues in predicting unplanned thirty-day readmissions. We address goal six by providing methods for measuring care variation in novel ways. Heart disease was cited by the report to be one of the top disease groups where improvements would have the most impact. In fact, diseases of the heart are the leading killer of Americans [5, 6]. As such, this research focuses on heart failure patients, as results found here may impact a leading health problem we face. In 2005, one out of every eight death certificates in America included

heart failure in the description for deaths, contributing to 292,214 deaths [6]. Similar death rates continue today [7]. It isn't that we aren't trying to address the problem, as pointed out in the WHO report, we are. In the 2010-11 fiscal year alone, the American Heart Association spent $110.9M on cardiovascular and stroke related disease research with $46.89M funding strictly cardiovascular disease research [8].

Such research often leads to new treatments for heart failure and other heart diseases, but the question remains: are we consistently employing current treatments in the ways that we should? WHO and the Institues of Medicine say that we are not [3,4]. But even if we are able to treat patients consistently, do we have adequate ways to know if they are ready to be discharged? This dissertation provides methods to address these important quality of care questions.

1.2 Problem Definition

Of great concern to many researchers in the medical community is the consistent application of the most up-to-date treatments across all patients who should receive them. Much research is done to measure consistency of treatment through care variation studies across patient groups aggregated across many care organizations and groups of patients. However, fewer methods are available to help care organizations understand the variation within their own organization. In addition, most care variation research is concerned with measuring variation of only a couple of treatments and ignore the remaining procedures used to treat a given condition. They also have no way of measuring the inherent variation of care within the population, only between specified populations. We provide means for care organizations to measure variation across many procedures used to address a given condition. We also provide a method for mea-

suring the variation in care within a patient cohort. These methods provide a way to measure the consistency in treatment both within and and across a care-organization's patient groups. However, even if we can measure and achieve consistent application of the most modern treatments, we must be sure that we do not release our patients when there is a high probability they will return sooner than is acceptable. Many researchers address this problem through modeling unplanned thirty-day readmissions.

Predicting unplanned thirty-day readmissions has three hierarchical problems to address. The first problem is the choice of data, or more specifically, variable selection: what variables are valuable and informative when predicting unplanned thirty-day readmissions. For practitioners hoping to reduce unplanned readmissions and thereby improve patients' care quality, knowing which factors may be related to unplanned readmissions is important. When resources are constrained, knowing which data is most predictive and being able to make informed, difficult choices is crucial.

One important issue is data availability. Most of the hospitals in the United States do not have advanced electronic health record systems and are therefore unable to implement many of the models proposed in the readmissions research. To address this issue, which so many care organizations face, we propose and assess models that use non-clinical variables from billing data. This data is found at all care organizations utilizing Medicare or Medicaid or most insurers.

Another major question asked by the research community is whether obtaining clinical variables would actually improve readmission prediction. Some research has suggested that clinical variables are not necessary for highly predictive models of readmissions. This dissertation will address both questions directly by comparing models with and without clinical variables. We also address an implied question raised by the first: which clinical variables and non-clinical variables are predictive of readmis-

sions.

Inherent in readmissions data is class imbalance. The majority of heart failure patients do not return to the hospital for unplanned visits [9]. This class imbalance may have implications for classifiers' performance. Little research dealing with this issue has been carried out in the readmissions literature. To properly improve patient care, it is important to know whether techniques for addressing imbalance are helpful and to what degree. This dissertation addresses class imbalance in medical data, specifically heart failure thirty-day readmissions data, by assessing the cost penalty method as well as three sampling techniques: over-sampling, under-sampling, and SMOTE.

Predictive modeling is woven through and is the means for assessing the issues just discussed. As such, the final matter is one of algorithm choice: which predictive algorithm is best for which scenario in predicting readmissions. Many current studies implement multivariate logistic regression when predicting readmissions despite advances in machine learning techniques. Those studies using advanced machine learning techniques have found little gain in predictive power. This dissertation compares the performance of logistic regression, Random Forests, and support vector machines in predicting unplanned readmission prediction.

New medical techniques continue to be developed but evidence has shown that we do not gain proportional ground to our investments [3]. Unless we have ways of measuring the consistency of their application, we cannot know if new techniques are being consistently applied across and within patient groups. This dissertation presents ways to measure that consistency both within and between patient cohorts. And even if those techniques are applied consistently to patients, it may be to no avail if those patients are released too early from the hospital. Predictive models capable of predicting the probability of an unplanned readmission allow for proper follow up to

better assure patient quality of care. This dissertation provides metrics and methods for improving quality of care both during patients' visits and just as they are released.

1.3 Hypotheses

The hypotheses addressing the issues mentioned above are arranged in four sections: Care Variation, Data, Class Imbalance, and Algorithm. We present the hypotheses and then the section where the results of the test are found. These hypotheses represent a significant part, but not all, of the contributions of this dissertation. The results of these hypotheses have lead to other contributions and recommendations. For reader convenience the table and section labels are hyperlinks when viewed electronically. Similar links are present in the Contributions section at the end of this dissertation linking to parts of the dissertation that address the described contribution.

1.3.1 Care Variation

H1.1 Distributions of procedures in response to each principal diagnosis are not significantly different across selected patient groups. Section 3.4.1

H1.2 Procedure distributions in response to a given primary diagnosis are not significantly different across physicians. Section 3.12

H1.3 Variation attributable to secondary and primary diagnoses is greater than that attributable to only primary diagnoses. Sections 3.4.1 & 3.4.2

H1.4 Measures of care variation do not significantly improve prediction of visit charge over a base model. Section 4.4

H1.5 No one measure of care variation significantly outperforms the others in improving prediction of visit charge over a base model. Section 4.4

1. Gini index
2. Entropy
3. Standard Deviation

H1.6 Measures of care variation do not significantly improve prediction of unplanned thirty-day readmissions over a base model. Section 6.4.1

H1.7 No one measure of care variation significantly outperforms the others in improving prediction of unplanned thirty-day readmissions over a base model. Section 6.4.1

1. Gini index
2. Entropy
3. Standard Deviation

1.3.2 Data

H2.1 Nonclinical data. Simple billing data does not significantly improve models' predictive performance over a base model of control variables. Section 5.4

1. Diagnosis count data does not significantly improve models' predictive performance over a base model of control variables. Section 5.4
2. Procedure count data does not significantly improve models' predictive performance over a base model of control variables. Section 5.4

3. Diagnosis count data combined with procedure count data does not significantly improve models' predictive performance over a base model of control variables. Section 5.4

H2.2 Clinical data. We hypothesize that the addition of explicit or implicit clinical variables in the form of laboratory test results does not significantly improve predictive performance over models without those variables. Chapter 6 starting at Section 6.4.2

1. Explicit - laboratory results. We hypothesize that no one laboratory result derived measure significantly outperforms any other. Section 6.4.2
   (a) Minimum lab result
   (b) Maximum lab result
   (c) Sum of squared abnormality
   (d) Lab result indicator variable
   (e) Last lab result
   (f) Number of times lab was run
2. Implicit - medications administered; We hypothesize that no one medication derived measure significantly outperforms any other. Section 6.4.2
   (a) Maximum number of times medications administered in a day
   (b) Total number of days each medication administered
   (c) Number of times each medication was administered
   (d) Gini of medication counts
   (e) Entropy of medication counts

(f) Standard Deviation of medication counts

H2.3 We hypothesize that laboratory result derived metrics do not significantly improve base models over medication derived metrics. Section 6.4.2

H2.4 We hypothesize that novel nonclinical metrics, thirty-day readmission indicator and number of prior inpatient visits, do not significantly improve performance of models. Section 6.4.3

1.3.3 Class Imbalance

H3.1 We hypothesize that schemes to address class imbalance will not significantly improve model performance over models built on non-adjusted data sets. Chapter 8

H3.2 We hypothesize that no one method addressing class imbalance will significantly outperform another. Chapter 8

1. Oversampling
2. Undersampling
3. SMOTE

1.3.4 Algorithms

H4.1 We hypothesize that no algorithm will significantly outperform another in predicting thirty-day readmissions. Chapters 5, 6, 7, & 8.

1. GLM: base vs. RF chosen variables. Chapter 8
2. RF: mtry, number of trees Chapters 5 & 6

3. SVM: linear vs RBF Chapter 7

1.4 Data

This research is based on data derived from the University of Virginia Clinical Database Repository (CDR) which is maintained by the Department of Public Health Sciences Clinical Informatics Division [10].

> The CDR is a frequently updated relational data warehouse that provides users with direct access to detailed, flexible, and rapid retrospective views of clinical, administrative, and financial patient data for the University of Virginia Health System. ... Its purpose is 'to meet the challenge of providing a way for anyone with a need to know — at every level of the organization — access to accurate and timely data necessary to support effective decision making, clinical research and process improvement'. [11]

The CDR houses records for over 1,000,000 patients spanning more than 15 years of patient data. While the CDR contains other forms of data, the data used for this paper was extracted from hospital billing records. Our data was generated using the following general conditions: *a*) Principal/Secondary diagnosis: Congestive heart failure (ICD-9 code 428.0 [12]) *b*) Date of diagnosis: between January 1, 2006 and December 31, 2010. 16,126 patients meet these criteria with a total of 62,892 total patient visits.

The CDR contains demographic information such as age, gender, ethnicity, as well as de-identified patient and case numbers, in- or out-patient status, length of stay for a given visit, year of admittance, visit hospital and physician charge, and data source. The dataset contains all the procedures a patient has been given, which are coded primarily in the Current Procedural Terminology [13] with some coded in International

Classification of Diseases, Ninth Revision (ICD-9) format [12]. All diagnoses are coded according to the ICD-9 standard.

When patients enter the UVA Health System they are assigned a permanent de-identified (i.e., random) patient ID number as well as unique visit identification numbers for each visit [14]. The care organization and physicians keep records of their assigned diagnoses and associated procedures for each visit [15]. The procedures, diagnoses, and charges are used to generate bills that are then transmitted the Centers for Medicare and Medicaid and most insurers [16] in electronic format.

1.5 Organization of the Dissertation

The remainder of this dissertation explores how care variation metrics and predictive modeling can be used to address current gaps in quality of care for patients during their visit and upon discharge.

Chapter 2 reviews the state of the art for the quality of care issues in terms of care variation and early unplanned readmissions. We start by describing care variation research and how it is currently carried out and how this dissertation addresses some of its shortcomings. We then describe the form of the method for measuring variation in care between patient cohorts. We then review the three measures for accounting for variation within a patient cohort. Chapters 3 and 4 address these gaps.

Chapter 3 presents a framework for thinking about the problem of care variation and then uses the χ^2 test using Monte Carlo simulation to test that variation across several defined patient cohorts. Chapter 4 explores the use of the within cohort metrics proposed in the literature review as predictors in simple models predicting visit charge. These metrics also appear in Chapters 5 through 7.

Chapter 2 continues with a review of current thirty-day readmission prediction for heart-failure patients. It explores the algorithms used and their performance. Chapter 5 explores the use of cost penalization in Random Forests as well as the usefulness of billing data in predicting readmissions. Chapter 6 compares Chapter 5's models to the same with the addition of clinical variables and then again with clinical variables plus two additional non-medical variables. Chapter 7 creates Support Vector machine models using the same data as Chapter 6. Chapter 8 compares sampling schemes used to address class imbalance across select datasets and the GLM, SVM, and Random Forest algorithms. Finally, Chapter 9 concludes the dissertation and discusses contributions and future work.

Chapter 2

Literature Review

In this chapter we review the state of the art for the quality of care issues in terms of care variation and early unplanned readmissions. We start by describing care variation research and how it is currently carried out and how this dissertation addresses some of its shortcomings. We then describe the form of the χ^2 test for measuring variation in care between patient cohorts. We then review the three measures accounting for variation within a patient cohort. Next is a review of current thirty-day readmission prediction approaches for heart-failure patients. We then review three of the algorithms used for modeling thirty-day readmissions and conclude with a section on methods commonly used to address class imbalance in predictive modeling.

2.1 Care Variation

One way researchers address quality of care issues is by measuring the consistency in patient treatment through care variation studies. Care variation is the difference in treatment patients receive given the same or very similar diagnoses. It is generally

accepted that variation for patients that could be considered to have the same conditions is undesirable. Another widely held belief is that for a given condition there is an optimal treatment. These two propositions together are what lead to the formation of treatment guidelines.

Guidelines are meant to represent the optimal treatment given a patient's condition. If guidelines represent the state of the art in treatment as they are derived from the most authoritative studies, then care variation can alternatively be defined in terms of variation from the optimal treatment as prescribed in guidelines. If variation is observed with reference to guidelines, then this would allow for comparison to a "gold standard" within many levels of care providers. In fact, many of the studies reviewed check for compliance to specific cases that are addressed in the guidelines.

Most studies of care variation use compliance rates on a specific patient population with regards to particular treatments as defined by guidelines. These are useful for understanding trends across a broad demographic but do not address localized problems such as doctor-to-doctor care variation within a clinic. These methods also do not describe variation across all procedures used to address a given diagnosis.

Other studies use physician surveys or other qualitative self-reporting methods, but the reports are known to not accurately represent what physicians actually do [17]. The psychology behind self-reporting leaves the data subject to recall bias [18]. Self-reporting also leaves out details that are useful in variation studies. In other words, physicians neither adequately describe their treatment choices and reasoning, nor do so in great detail. Another pitfall of self-reporting is that it interrupts the physicians and prevents them from practicing medicine. This fact makes such methods logistically difficult for the purpose of monitoring change in care variation within a care provider organization.

In order to measure physician practice at a provider organization, variation studies specific to that organization are needed. The large studies set up to observe variation are generalized across large populations and are unable to observe local variation. Qualitative surveys and other interaction based methods interrupt care and are intractable for providing the detail needed to inform decision makers on change in care variation. Prior approaches to studying care variation are neither set up to monitor individual provider organizations nor are capable of measuring variation at various care provider levels.

The problem is that there are inadequate methods to measure care variation across many procedures used simultaneously to address a specific condition both between providers and within provider organizations. While much research on care variation across specific populations of patients has been performed we still lack methods to quantify and monitor care variation for many conditions within and across levels of care (such as clinics, departments, hospitals, and entire health systems). Table 2.1 summarizes the next sections that review the state of the art in care variation research.

2.1.1 Common Approaches

Many have characterized care variation for various medical conditions including heart failure. The most prevalent approach expresses care variation by percentage or odds ratio of a constrained population using specifc treatments [19–37]. These studies look at particular patient populations and assess the rate of compliance or odds of using certain recommended treatments for that condition. The odds ratios are derived from modeling care variation using (multiple) logistic regression for classifying patients into treatment groups [38–45] and comparing those groups for consistent application of

appropriate treatments. Groupings can include geographic [25, 29], racial/ethnic [45], age [21], and others. Others use multilevel modeling [46–48] to model variation effects. Jaglal et al. and others compare the rates of compliance across time [49–51] while Boarj and Gallerani also detect seasonality for various timespans [52, 53].

2.1.2 Qualitative Methods

Not all approaches are primarily quantitative. Reis et al. [54] compare the treatments chosen by a physician generalist to those of a cardiologist using chart reviews of 160 patients, finding differences in treatment and outcome between the groups. Chart review is logistically difficult to scale to thousands or millions of patients. Some researchers [55–60] use interaction based methods such as surveys or group sessions to develop understanding of doctors' care variation decisions and rationale. Such approaches are difficult to orchestrate and quickly scale to thousands or millions of patients although this may be possible with sophisticated organization and logistics. Another issue with interaction based methods is that such methods are prone to recall bias [18, 61]. For example, Sboner and Aliferis [17] model clinician's treatment choices and why they make them, and compare those judgements with self-reported justifications using support vector machines and markov blanket variable selection. The authors find that the physician's choices do not reflect the guidelines they claim to follow, which is strong evidence of recall bias. Such bias makes the interaction based methods less reliable as a data source. Another simple, but important, issue with interactive based methods is the interruption to physicians' work, which stops them from actually treating patients. This sort of time drain aggregated across the nation could have costly effects.

2.1.3 Non-localized Studies

Data registries containing thousands and millions of patients' data with hundreds of contributing hospitals have enabled many of the care variation studies already mentioned. Using registries for studying variation is common in many fields including heart disease, lung cancer, renal failure, and many others [19, 37, 62–64]. Example registries include ADHERE [65], Optimize_HF [66], and SOLVD [67]. While useful for generalized statements of care variation, the registry based studies are general to many hospitals and specific to certain diagnoses and patient types. For monitoring and practical, local change, care variation should be measured for a specific care provider organization (such as a hospital) and be able to include many diagnoses and patient types.

2.1.4 Query Driven Methods

Some studies do move beyond the simple comparison of treatment compliance rates and use more sophisticated methods for extracting treatment patterns but are not able to describe them. For example, Wang et al. [68] identify complex treatment and diagnosis patterns using a visual tool. While useful, the tool lacks generalized theoretical underpinnings and analysis to understand the underlying patterns beyond visualization making measuring, monitoring, and comparing variation nearly impossible. Similarly, Plasaint et al. [69] develop an interactive query tool to help practitioners find "specific temporal patterns in both numerical and categorical data" for radiology patients. These tools help uncover more complex, albeit directly queried, patterns but not latent, abstract, or general patterns. Because these patterns are query driven and do not have associated methods or theory for characterizing them for later analysis, it is

difficult to generalize across many patients and monitor and compare for the purpose of understanding variation.

2.1.5 Generic Pattern Detection

Much of the work that moves beyond simple rate comparisons does detect patterns but does not directly measure care variation [70–73]. Huang et al. [74] use decision tree induction to find chronic disease rules and then apply association rules to frequent item sets generated by the decision trees to associate disease cases. Patil and Kumaraswamy [75] develop a heart attack prediction system using K-means clustering to find relevant data to heart attacks and then the MAFIA algorithm to detect patterns. Huang et al. [76] use association rules to identify common comorbidities for patients with Obstructive Sleep Apnea. Li et al. [77] find risk patterns, using optimal rule discovery algorithms. They focus on very rare, abnormal events, and do not characterize the entire distribution of outcomes. Ryan et al. [78] found symptom clusters for patients with acute myocardial infarction using latent class analysis [79] which is related to latent semantic analysis. They identify groups of patients based on some latent variable extracted or defined by the method. Ting et al. [80] develop an automatic Medical Knowledge Elicitation System (MediKES) to capture physicians' tacit knowledge in a machine readable form. These methods offer insight into how patterns may be discovered and modeled but do not directly address care variation, but with further work many of these approaches could be amended to address it.

2.1.6 Summary

The majority of care variation studies compare rates or odds ratios of treatment usage over specific patient populations. Many of these studies are based on data registry patients and are generalized over hundreds of hospital service areas making them non-localized and difficult for the use of local variation monitoring. Interactive methods are amenable to localized study but are subject to recall bias and distract from patient care. Some that do go beyond the simple rate comparisons are query driven and thus do not allow for consistent measurement of care variation nor are they generalized, discovered patterns. Some methods do discover patterns using modern data mining techniques but do not directly assess care variation. No single method is locally applicable yet simultaneously applicable to many diagnoses, detects generic patterns, allows for monitoring, and does not distract from patient care.

Table 2.1: Summary of past approaches to characterizing care variation and treatment patterns.

§	Approach Summary	Papers	Remarks
2.1.1	Care variation by percentage or odds ratio of a population using specified treatment(s)	[19–37, 201, 202]	Limited to specific populations, diagnoses, and treatments
	Logistic regression models of treatment variation for various diseases including heart failure.	[38–45]	”
	Multilevel modeling of variation effects	[46–48]	”
	Compare the rates of compliance across time	[49–51]	”
	Detect seasonality for various timespans	[52, 53]	”
2.1.2	Chart review comparing treatments & outcome of generalist to specialist	[54]	Difficult logistically to scale

Continued on next page

Table 2.1 – *Continued from previous page*

§	Approach Summary	Papers	Remarks
	Interaction based methods: surveys, group sessions	[17, 55–60]	Recall bias present, interrupts care
2.1.3	Data registries for aggregate outcome measures	[19, 37, 62–67].	Summaries over large populations of specific diseases and treatments, non-localized
2.1.4	Visualization or query tool to identify complex patterns	[68, 69]	Limited to user query and visualization, not generalizable
2.1.5	Prediction/Classification of treatment courses	[70–73]	Prediction modeling of treatment, doesn't characterize variation
	Decision tree induction creating disease rules, association rules to assign disease cases	[74]	Classification, not care variation
	K-means clustering to filter data, MAFIA algorithm to predict heart attacks	[75]	Outcome prediction, not care variation
	Association rules identify common patient comorbidities for sleep apnea	[76]	Addresses complex relationships but not care variation
	Optimal rule discovery to find outlier risk events in patients	[77]	Focus on extreme events, not patterns in care variation
	Latent class analysis to cluster heart attack patient sub-groups	[78]	Not care variation
	Text mining EHR to create physician machine readable form representing tacit physician knowledge	[80]	Models physician decision based on EHR notes

2.2 The χ^2 Test

The χ^2 test for independence is the proposed metric for measuring care variation between patient cohorts. We describe its development and issues regarding its appropriate use in the following sections.

2.2.1 Development and Definition

The X^2 test was developed by Pearson in 1900 [81] to test goodness-of-fit between a vector of values and a distribution of choice. For years following Pearson's initial X^2 test, scholars debated on appropriate degrees of freedom and the asymptotic properties of his proposed test [82]. Fisher [83] refined the test and described his exact χ^2 test in 1922 resolving some of the issues. We describe the test here following Cochran's derivation in his well-known 1952 paper [82].

The standard test has several assumptions and definitions:

1. n observations form a simple random sample from a population.
2. The observations fall into mutually exclusive classes.
3. p_i is the probability an observation falls into the ith class, $i = 1, \ldots, k$.
4. The p_i are specified by either a theory of known numbers or parameterized functions.
5. $m_i = np_i$ are the expected cell values, where
   (a) $\sum_{i=1}^{k} p_i = 1$ and
   (b) $\sum_{i=1}^{k} m_i = n$.

Given these assumptions and definitions we define describe the joint distribution of the observations, x_i being in each class with the multinomial distribution:

$$\frac{n!}{x_1!x_2!\cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}. \tag{2.1}$$

To test whether the joint distribution of the observations was truly multinomial, Pearson proposed the X^2 criterion formulated as

$$X^2 = \sum_{i=1}^{k} \frac{(x_1 - m_i)^2}{m_i} = \sum_{i=1}^{k} \frac{(x_i)^2}{m_i} - n \tag{2.2}$$

where there was no stated alternative hypothesis.

The limiting distribution of (2.2) when $n \longrightarrow \infty$, with fixed p_i is the χ^2 distribution,

$$\frac{1}{2^{v/2}(\frac{v}{2}-1)!}(\chi^2)^{v/2-1e^{-\frac{1}{2}\chi^2}} d\chi^2 \tag{2.3}$$

where v is the number of degrees of freedom. If we assume $n \longrightarrow \infty$ and also assume fixed p_i this implies that our $m_i \longrightarrow \infty$. This means that our cell values must be large in order to use the χ^2 distribution to test our hypothesis. To address the need for large m_i many proposed that the minimum expected cell values be at least 5, some saying as high as 10. These numbers are really only tradition and Cochran suggests in some circumstances it could be as low as 2 when 30 degrees of freedom are present. Many have contributed to corrections and adjustments [84–90] and supply ample suggestions for a myriad of situations. An example solution to this quandary is the combination of classes, which may be done arbitrarily, though typically it is suggested that an expert does so to maximize interpretability of each class. The problem with this, according to Cochran, is a loss of power because adhering to this rule typically requires combining classes that occur in the tails or extremes of the distribution where differences tend to be most obvious. All of these issues may be overcome, however, given enough computational power as demonstrated in the next section.

2.2.2 Monte Carlo Hypothesis Testing

Monte Carlo significance tests had been proposed as early as 1957 by Dwass [91] and independently again in 1963 by Barnard during a discussion of a paper by Prof. Bartlett

[92]. In the discussion after Prof. Bartlett's paper presentation [93], Prof. Barnard responds to another's suggestion that an exact test isn't important in the case discussed, but then goes on to describe that one could calculate the exact probability of the statistic at hand if so desired. To do so, he says one could create a simulation using 19 data sets sampled from the null (one hypothesized to be the same as the one at hand) then calculate the statistic of interest on each and rank the now twenty statistics: 1 original + 19 simulated = 20 statistics total. One now has the exact probability of randomly obtaining a statistic greater than or equal to the original.

The idea was more fully described by Hope in 1968 [94] when she formally described its implementation and showed some of its properties in limited situations. The process is relatively simple and may be applied to any distribution, but for our purposes we use it for the χ^2 test.

The process for the nonparametric version as described by Hall and Titterington [95] is:

1. Obtain a sample X of size n from a population π.
2. Calculate the statistic of interest, T, from X.
3. Draw B samples with replacement of size n, $X_1^*, \ldots, X_N^*$, from X.
4. Compute T_i^* from X_i^*, just as T was calculated from X.
5. Rank $T_1^* \ldots T_N^*$ as $T_{(1)}^* \leq \cdots \leq T_{(N)}^*$
6. Reject H_0 if $T \geq T_{(M)}^*$, where $1 \leq M \leq N$.

The test's level is determined by choosing the appropriate M. The nominal level of the test is $\alpha = (N + 1 - M)/(N + 1)$ while the exact level of the test is $\alpha' = Pr(T \geq T_{(M)}^*)$ or the simple probability that T exceeds $T_{(M)}^*$. The p-value is the proportion

of $T^*_{(i)} \geq T$. The calculation of the Monte Carlo p-value is not without scrutiny but the aforementioned calculation is the accepted implementation. Recently, North et al. [96] suggested some corrections, but were quickly refuted with consensus around the conventional calculation proposed in the original formulation [97].

When the method was originally proposed, and for years after, computational power was limited, so issues of picking the appropriate number of simulations were important. Marriott suggested at least 100 simulations for when $\alpha = 0.05$ but stated that more is always better. Applications of the approach followed soon after its publication. Besag and Diggle [98] use Monte Carlo hypothesis testing for spatial patterns including bird migration pattern transference, contagion outbreak, and kittiwake nesting patterns. More modern papers have used the approach to test DNA sequences [99], ecological models [100], and even extend to sequential Monte Carlo hypothesis tests for big data [101, 102]. Besag and Clifford [103] generalized Monte Carlo hypothesis tests to situations where observation independence is violated. Friedman [104] discusses using Monte Carlo goodness-of-fit and two-sample testing with multivariate distributions.

The power of Monte Carlo tests is explored by Joeckel [105] and by Hall and Titterington [95]. Joeckel was able to show the power loss of using Monte Carlo (MC) hypothesis tests when an exact test is available as measured by Dwass efficiency. Where the exact test isn't known he showed the MC power loss using asymptotic relative Pitman efficiency (ARPE) and local asymptotic relative Pitman efficiency (LARPE). For example, if the exact test is known but MC tests are used for a sample size 999 and α is 0.05, then the MC power is 94.5% of the original test. If $\alpha = 0.025$, then the power is 92.1% of the original. For unknown tests of same size samples and confidence levels, the LARPEs are 99.5% and 99.4% of the original respectively. Hall and Titterington [95]

showed if the asymptotic distribution of our statistic of interest does not depend on any nuisance parameters, then the MC test has improved accuracy over asymptotic methods by an order of magnitude holding confidence level constant.

The Monte Carlo χ^2 test for independence is found in the `chisq.test()` function in the `stats` package in `R` [106] and is used by setting the option `simulate.p.value` to `TRUE`.

Monte Carlo hypothesis tests have been a viable alternative to exact and asymptotic hypothesis tests for many years. Given the computational power available in standard computers and ease of implementation, they are feasible for many research situations. Their properties have been explored including their minimal loss of power and even improved power in certain circumstances. They are well-suited to comparing empirical distributions of counts for assessing care variation in various patient cohorts.

2.2.3 Multiple Hypothesis Testing

Controlling Type I error (rejecting the null when it is actually true) is a concern when many hypothesis tests are performed in a single study. Type I error is controlled by choosing α and is called comparisonwise error or also the individual error. This implies that the probability of not rejecting a null hypothesis is $1 - \alpha$ when it is actually true. When multiple tests are performed, say k, then the error rate across the battery of tests is $(1 - \alpha)^k$. If we assume all k tests are independent and all null hypotheses are true, then the probability of rejecting at least one of the k null hypotheses is $1 - (1 - \alpha)^k$. This is called the familywise error rate. If we were to perform 2,383 such tests with $\alpha = 0.05$ then this would yield a familywise error rate of essentially 1. This means when performing so many tests, it's almost a given that you'll reject at least one null

hypothesis even when it is actually true.

Methods to control the maximum experimentwise error rate (or familywise error rate in the strong sense) are available, with the Bonferroni correction being the most highly recognized but also most conservative. It corrects by multiplying each p-value by m, the number of comparisons made, or equivalently, the familywise error rate becomes $1-(1-\alpha)^{1/m}$ or approximately α/m. Alternative corrections are available.

The Holm-Bonferroni correction was designed to completely replace the Bonferroni method. It is provably more powerful and is less conservative [107]. A close competitor to Holm's correction is the Hochberg correction [108] which requires independence between the tests whereas Holm's correction requires no assumptions be met. Holm's procedure is a sequentially rejective multiple test procedure and is as follows. First, choose α. Let there be n hypotheses $H_1, H_2, \ldots, H_n$ with corresponding test statistics $Y_1, Y_2, \ldots, Y_n$ with obtained p-values $P_1, P_2, \ldots, P_n$. We then reorder the p-values from smallest to largest denoted by $P^{(1)} \leq P^{(2)} \leq \ldots \leq P^{(n)}$ with corresponding hypotheses $H^{(1)}, H^{(2)}, \ldots, H^{(n)}$. We then compare the reordered p-values to the numbers $\frac{\alpha}{n}, \frac{\alpha}{n-1}, \ldots, \frac{\alpha}{1}$ in that order. The first reordered p-value we fail to reject using the new numbers compared to our originally chosen α signals that we fail to reject all hypotheses from that hypothesis on.

The decision to control for familywise error rate is not always straight forward and is highly debated. Some propose that no adjustments should be made at all [109, 110]. The arguments against it are several. One objection is that the outcome of an experiment should not be dependent on the number of other tests being performed. To quote Perneger [109]:

> In a clinical setting, a patient's packed cell volume might be abnormally low, except if the doctor also ordered a platelet count, in which case it could be

> deemed normal. Surely this is absurd, at least within the current scientific paradigm. Evidence in data is what the data say–other considerations, such as how many other tests were performed, are irrelevant.

Perneger's objections notwithstanding, it is generally accepted that one should adjust for Type I error when performing many tests, especially for clinical trials [111–113]. Bender and Lange [114] take the middle argument. They propose that when performing exploratory work, you should not make adjustments as you don't want to preliminarily close doors that could lead to great discoveries. However, when the work is confirmatory, then adjustments should be made [114]. As this is a decision complicated by many factors, the debate will surely continue.

2.3 Measures of Variation within a Distribution

The variation metrics have only one property to satisfy: higher values represent less uniformity and conversely, greater variation. Two of the three are normalized between zero and one (Gini and Normalized Entropy) while standard deviation is not.

The first variation metric under consideration is the standard deviation of patients' average daily procedure application rates, $SD(X)$. The standard deviation is that typically used in statistics,

$$\sigma = \sqrt{E[(X - E[X])^2]}. \tag{2.4}$$

This measure of variation is probably most familiar and provides a benchmark visit complexity metric to which developed metrics may be compared. It uses the mean as the reference of deviation. It squares deviations from the arithmetic mean of the

distributions. This means that the standard deviation weights larger differences from the mean more than it weights smaller differences from the mean.

The second metric is the Gini coefficient [115], $G(X)$, which is used to measure dispersion and is applied in Economics to measure income inequality. Mathematically, it is defined as

$$G = 1 - \frac{\Sigma_{i=1}^{n} f(y_i)(X_{i-1} + X_i)}{X_n} \tag{2.5}$$

where $X_i = \Sigma_{j=1}^{i} f(y_j)\, y_j$ and $X_0 = 0$. Its range is 0 to 1 with 1 having the greatest dispersion and 0 being uniformly distributed.

The Gini coefficient measures deviation using a different mechanism. It is mathematically equivalent to one-half the relative mean difference of the distribution [116]. This means that the Gini coefficient does not compare to a central statistic. Rather, it compares the distribution's entries to themselves and provides a summary of the mean difference in randomly chosen pairs from the distribution. Thus, the Gini coefficient does not weight larger distribution values like the standard deviation does.

The final metric is a variant of entropy. High entropy means more uniformity in distribution. This means that lower entropy means greater variation in a distribution. To keep our interpretation of the complexity measure consistent we normalize the entropy by dividing by $\ln N$ (a vector's maximum entropy), where N is the length of the vector's nonzero entries, and subtract from one so that higher normalized entropy corresponds to higher complexity. If X has $i = 1 \ldots N$ observations each denoted x_i then our normalized entropy measure is thus defined as

$$NE(X) = 1 + \frac{\sum_{i=1}^{N} p(x_i) \ln p(x_i)}{\ln N}. \tag{2.6}$$

Normalized entropy as a variation measure takes a different approach altogether.

It looks at each distribution entry as a state and measures the overall uniqueness, or alternatively the uniformity, of all the states in the distribution. However, this uniqueness is much more heavily weighted toward unique states that are larger in value than the remaining entries. This results directly from entropy's formulation. A single large value multiplied by its natural log added to the summation increases entropy much more than a small value multiplied by its natural log that is then added to the summation.

2.4 Thirty Day Readmissions Prediction

Thirty day-readmissions prediction is of general interest for many reasons. One reason stems from the adoption of the Patient Protection and Affordable Care Act in 2010 and more particularly its hospital readmission reduction program [117, 118]. This program will penalize hospitals who have not met readmission performance standards across five disease types of which heart failure is included. Readmissions prediction research predates this act, of course, as it may be a useful metric for hospitals to simultaneously assess their quality and profitability. Its usefulness as a metric is much debated [119–123]. Despite the debate, the metric is set at a legislative level and will continue to be important for many years. Predicting thirty day readmissions accurately is therefore an important problem that has both quality of care and fiscal consequences. Further research is needed in developing variables and finding appropriate models to accurately predict readmissions.

2.4.1 Variable Choices

Some thirty day readmission studies rely only on administrative data found in hospital billing systems. This data is already collected for billing purposes and is used in after-the-fact observational studies for modeling readmissions. Because this data is standardized and used for hospital billing, regulatory reporting, and utilization tracking, any models based on it will be widely applicable. For example, the Centers for Medicare & Medicaid Services' (CMS) report and others [124–126] used administrative hospitalization data at fee for service facilities for predicting readmissions.

As more advanced EHR systems come online and clinically meaningful data becomes available for analysis, both real-time and retrospective thirty-day readmission studies include them in their models. For example, Amarasingham et al. [127] incorporate 17 laboratory and vital sign variables into their automated prediction model. While some have shown that clinical variables did not improve prediction [128, 129] Au et al. demonstrated a marked improvement in readmission prediction [130]. Bradley et al. demonstrate the ability of the Rothman Index [131], which incorporates 26 clinical variables and vital signs, to predict readmissions [132].

Two patterns weave through the aforementioned studies. First, data is often aggregated into single variables. The LACE and LACE+ scores, Tabak score [127], and Rothmann Index [132] are examples of clinical and administrative variables being combined to form indices. This is often done as single numbers are easy to use for making clinical decisions [125]. Their merits notwithstanding, aggregating information into a single number inherently reduces the information available for accurate prediction. Moreover, many indices use simple additive (e.g., Charlson index) or other functions which may be improved by using simple linear regression or other methods that are designed to optimally weigh variables to model outcomes such as the readmissions.

The second pattern is explicit variable choice by physicians when creating models. While this makes intuitive sense (many physicians have practiced medicine for years), it does leave out variables which may have presently unknown relationships to readmissions. Both of these patterns, aggregating variables, and physician variable selection have counterparts in machine learning: model selection, and variable selection. In machine learning, these counterparts may be optimized, validated, and measured objectively. This is the heart of data-driven analysis.

While clinical variables are useful, they are not widely available as only 8.7% have the most advanced level of EHR systems according to the Healthcare Information Management and Systems Society, a EHR standard setting authority [133]. Another issue with detailed clinical data is that even in modern EHR systems the data representation is not fully standardized between sites. Thus, methods developed in some advanced systems may not be useful in other advanced EHR systems. However, basic medical data in the form of standardized billing data exists for any hospital using Medicare or Medicaid (and most private insurers) due to their reporting requirements [16]. Worse yet, those hospitals with less advanced systems tend to be small and rural [133]. It is the poor, rural population that tends to have increased rates of heart disease [134] Therefore, the populations who are in greatest need for improved healthcare regarding heart disease are those that do not benefit from studies using advanced EHR data. To address this issue, in part of this research, we use administrative data to allow the results to be applicable to any hospital in the United States and elsewhere.

2.4.2 Algorithms Utilized in the Readmissions and Medical Literature

Many readmission studies use more traditional statistical modeling techniques such as logistic regression [124, 126, 131, 135] and Cox proportional hazard models [136–143] in their readmission research [125, 144].

Advanced machine learning techniques have been used in healthcare informatics. Fonarow et al. used regression and classification trees for stratifying heart failure patients into risk groups for in-hospital mortality [125]. Support Vector Machines have been used to identify predictors of medication compliance in heart failure patients [145]. Random Forests analysis has been performed for predicting readmissions based on claims data combined with advanced EHR data such as vital signs, medication orders, and others [130, 146, 147]. A variant of random forests, random survival forests, has been used to analyze heart failure survival using laboratory results and vital sign data such as stress test results [144].

Table 2.2 gives a review of models and their performance when used to predict heart failure thirty-day readmissions. The metric used for model performance is area under the receiver operating characteristic (ROC) curve or AUC. This metric is also called the c-statistic [132] and is the most often used in the thirty-day readmissions prediction literature [148]. Though some find issues with the metric [149] it is the most commonly accepted metric for choosing classifiers and best demonstrates the overall performance tradeoff in imbalanced data [150, 151]. For medical diagnostics problems, it has been found to be the best overall single number metric for choosing between machine learning algorithms [152]. For these reasons we will use it exclusively throughout this dissertation. In the table, it is clear that the vast majority of research in this area

has utilized logistic regression as the technique of choice with very little exploration outside of that norm. What other models had been used did not appear to improve performance which may partially explain why more research has not utilized machine learning techniques. Only two models on general heart failure patients (not subdivided into medical vs. surgical) have AUCs greater than 0.7 and they are Bradley's Rothman Index and Amarasingham's electronic readmissions model. None of the models are even close to 0.8 which is generally accepted as the threshold for good prediction.

Table 2.3 shows a summary of each modeling group's AUCs as well as the all together. The best AUC overall was 0.78 but isn't as comparable to the remaining AUCs as it was on a very specific subgroup of patients, those who had undergone surgery. The second highest AUC was 0.73, the counterpart to the 0.78 model with those patients who were treated both medically and surgically. The next best, at 0.72 AUC, came from the well-cited Amarasingham et al. paper. The apparent gap between modern machine learning techniques and performance of modern readmissions prediction models is quite large. The worst performing model was just above random at 0.54 and was a Random Forests (RF) model. The median overall was 0.63. The mean for the GLM models was 0.64 and 0.62 for the other models. The underperformance of the advanced machine learning models may have something to do with the lack of followup research utilizing the algorithms but this is conjecture.

Overall the performance across the general heart failure population models is poor (as opposed to the more specialized exceptions already noted). Kansagara et al. [148] pointed out the same pattern across the general readmission literature (which included heart failure readmissions). This dissertation aims to address this lack in performance and provide new variables, models, and approaches that are directly applicable to heart failure patients, and plausibly to many others.

Table 2.2: Performance of readmissions models in the literature as measured by AUC.

Algorithm	Model	AUC	Paper
Logistic Regression			
	Krumholz	0.60	[153]
	Hammill Claims Only	0.59	[128]
	Hammill Clinical & Claims	0.60	[128]
	Keenan Administrative	0.60	[126]
	Simple Scoring System	0.60	[154]
	Weighted Scoring System	0.61	[154]
	Rothman Index (RI)	0.62	[132]
	Meadem	0.64	[155]
	Bradley's RI General	0.73	[131]
	Bradley's RI Medical	0.72	[131]
	Bradley's RI Surgery	0.78	[131]
	Gildersleeve	0.70	[156]
	ADHERE	0.56	[127]
	CMS Risk Adjustment	0.66	[127]
	Tabak	0.61	[127]
	Electronic Readmissions Model	0.72	[127]
	Zolfaghar LR Original	0.64	[157]
	Zolfaghar LR Oversampling	0.63	[157]
	Zolfaghar's Yale	0.59	[157]
	Agrawal's Orig.	0.64	[158]
	Agrawal's Under	0.63	[158]
	Agrawal's Over	0.64	[158]
Random Forests			
	Charlson	0.54	[130]
	Au's Krumholz	0.58	[130]
	Au's Keenan	0.58	[130]
	LACE	0.68	[130]
	LaCE	0.60	[130]
	Zolfaghar RF Original	0.61	[157]
	Zolfaghar RF Oversampling.	0.62	[157]
Support Vector Machines			
	Meadem et al.	0.64	[155]
	Agrawal's Orig.	0.63	[158]
	Agrawal's Under	0.64	[158]
	Agrawal's Over	0.63	[158]
Naive Bayes			
	Meadem et al.	0.64	[155]
	Agrawal's Orig.	0.65	[158]
	Agrawal's Under	0.63	[158]
	Agrawal's Over	0.60	[158]

Table 2.3: Performance summary of readmissions models in the literature as measured by AUC.

	GLM	Other	All
Min.	0.56	0.54	0.54
1st Qu.	0.60	0.60	0.60
Median	0.63	0.63	0.63
Mean	0.64	0.62	0.63
3rd Qu.	0.66	0.64	0.64
Max.	0.78	0.68	0.78

2.5 Classification Algorithms

2.5.1 Logistic Regression

Logistic regression is a a linear classification method. It can be used for classifying multiple classes but for our purposes it will be used for two classes.

Logistic regression uses the logit link function to transform the expected value of the response to the log odds of the expected value. The logit link function is $\log[p/(1-p)]$ resulting in the logistic regression function

$$\log \frac{Pr(G=1|X=x)}{Pr(G=2|X=x)} = \beta_0 + \beta^T x. \tag{2.7}$$

The boundary between classes, or decision boundary and is the set of points where the log-odds are equal to zero [150]. The coefficients are fit using iteratively reweighted least squares using deviance as a loss function. Although fast, especially with modern computers, the solution for the coefficients is not closed-form and so may take a while to compute and may not even converge. Because this is a generalized linear model the coefficients have a direct input-output interpretation. This is especially useful to practitioners who wish to understand the relationship between the features and the

response. The `stats` package in `R` has a function `glm` that is used in this research.

2.5.2 Random Forests

Random Forests are an example of ensemble methods (also, Random Forests is trademarked by Salford Systems and is thus capitalized throughout this work). Random Forests utilize many classification trees to average out their variance to achieve a prediction. The method creates many decision trees (many trees together make a forest, thus the use of forest in the name) and at each node in a tree it randomly selects $\sqrt{p}$ (where p is the number of variables in the dataset) variables to choose between for selecting the best variable and variable value simultaneously. The choice is based on an impurity metric such as the gini coefficient. It then continues down the decision tree. The tree then uses the majority vote of the end nodes to determine the class of the observations. These votes are then summed across the forest of trees to determine the predictions for the observations. These predictions are given using a probability (really, a proportion of trees or even nodes, that voted one way or the other).

Choosing variables randomly to consider at each node is important as it keeps the trees independent. As the number of variables chosen at each node increases, this independence between trees is compromised and the generalizability of the model is reduced. If more variables are needed at each node this is usually an indication of noisy variables which do not contain much information and therefore more are needed to reduce the variance in the prediction [159].

Variable Importance

One desirable feature of having the trees independent, is that it makes the algorithm embarrassingly parallel. This means that trees may be grown on different computing nodes and then recombined later to produce the forest and its predictions. Because Random Forests can be a relatively fast procedure and can be used on "wide" problems it allows us be very creative in our variable creation and selection.

One valuable output of the Random Forests algorithm is variable importance. Variable importance is measured by the so called 'mean decrease in accuracy' is the average difference in out-of-bag (OOB) error when the variable in question is permuted from its original value in a given tree [160].

The following description of calculating variable importance follow's Genuer et al.'s excellent version [160] closely. The variable importance of a given variable, X_j is defined as:

1. For each tree, t
   (a) Calculate $errOOB_t$, the MSE in an out-of-bag (OOB) sample for tree t
   (b) Obtain $\widetilde{OOB}_{tj}$, by randomly permuting the values of X_j in tree t
   (c) Using $\widetilde{OOB}_{tj}$ calculate the error of tree t on the permuted sample, $\tilde{errOOB}_{tj}$
2. Calculate the variable importance of X_j as

$$VI(X_j) = \frac{1}{B}\sum_t (\widetilde{errOOB}_{tj} - errOOB_t), \tag{2.8}$$

   where B is the number of trees in the random forest.

The mean is calculated from the same operation on all the trees [161]. This provides

an importance ranking as it influences the predictiveness of the algorithm itself. While making no causal inference, it does provide a measure of importance (thus the name) for variables' ability to influence the predictiveness of the algorithm. It has also been shown that the variable importance found by Random Forests agrees with that found by linear regression. [162]. Mean decrease in accuracy has been shown to be more reliable than a second measure using gini importance which has been shown to place more weight on categorical variables with many levels [150, 163]. However, mean decrease in accuracy has been shown to unduly assign higher rank to highly correlated variables [164]. We note this limitation and accept it as a tradeoff for this implementation of random forests' computational efficiency.

2.5.3 Support Vector Machines

This section follows much of the derivation from *Elements of Statistical Learning* by Hastie et al. [150]. A support vector classifier is a linear classifier. The classifier becomes a machine when it utilizes various basis transformations on the data to achieve linear classification in higher dimensional spaces [150].

Support vector classifiers extend the idea of a linear classifier by adding the concept of the margin between two classes. The initial form of this algorithm, introduced by Vapnik [165], assumes that the two classes are perfectly separable. The idea is to find a hyperplane that separates the two classes while simultaneously maximizing the distance to the closest points that belong in each class and hyperplanes passing through these closest points, called the margins. For convenience the classes are denoted using

-1 and 1. The optimization problem is

$$\begin{aligned} &\max_{\beta,\beta_0,\|\beta\|=1} \quad M \\ &\text{subject to} \quad y_i(x_i^T\beta+\beta_0) \geq M,\ i=1,\ldots,N. \end{aligned} \tag{2.9}$$

The primal can be written

$$L_P = \frac{1}{2}\|\beta\|^2 - \sum_{i=1}^{N}\alpha_i[y_i(x_i^T\beta+\beta_0)-1] \tag{2.10}$$

with associated dual

$$\begin{aligned} L_D =& \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{N}\alpha_i\alpha_k y_i y_k x_i^T x_k \\ &\text{subject to } \alpha_i \geq 0 \text{and} \sum_{i=1}^{N}\alpha_i y_i = 0. \end{aligned} \tag{2.11}$$

In this formulation it is possible to have new observations fall inside the margin. If this occurs, then the separation may be optimal nominally, but perform poorly due to overfitting. For this reason the classifier is generalized using kernels and cost penalty to allow for some misclassification when solving for the support vectors.

Transforming the feature space makes the support vector classifier into a support vector machine, as different kernels for generating the transformed variables are interchangeable. We can choose a set of basis functions $h_m(x), m = 1,\ldots,N$ and then compute the new input features accordingly. This allows what would require a nonlinear boundary in the original space to now have a linear boundary in the transformed space. The dual including these functions can be written

$$L_D = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{N}\alpha_i\alpha_{i'}y_i y_{i'}\langle h(x_i), h(x_{i'})\rangle. \tag{2.12}$$

Support vector machines do not have a nice interpretable output like the prior two methods. The support vectors are built at the boundary of the classes and so do not provide a look at the entire feature space. This poses issues when interpretation is important to those using the algorithm.

2.6 Methods Addressing Class Imbalance

Class imbalance may have important implications for classifier performance. Approaches to address classification come in three general ways, sampling, algorithm modification, and cost sensitive learning [151]. López et al. [151] as well as He and Garcia [166] provide excellent reviews of empirical issues and insights when dealing with imbalanced data.

2.6.1 Cost Sensitive Learning

Cost sensitive learning applies a higher cost to one group than another when misclassified by the classifier using various algorithms and implementations. Many algorithms have costs built in, such as Random Forests, and support vector machines, while others have developed generalized costing methods [167, 168]. Breiman [159] claims that the cost adjustment in his implementation of Random Forests overcomes class imbalance problems. Wallace and Dahabreh have shown that many costing methods "systematically underestimate the probabilities for minority class instances" even when well calibrated [169]. Despite SVMs' built in costing, Tang et al. [170] were able to develop novel undersampling techniques to better address highly imbalanced datasets. Roumani et al. [171] tested several algorithms including SVMs and logistic regression on highly imbalanced ICU data using cost penalization methods and found that both

logistic regression and SVMs perform well. Not many heart failure readmissions studies have used sampling techniques. We mention those that have in each appropriate section.

2.6.2 Under- and Oversampling

Under- and oversampling methods seek to address the question posed by Chawla et al. [172]: "What is the correct distribution for a learning algorithm?" Sampling techniques change the actual class representation seen by the learner to equalize their presentation in the dataset.

Undersampling

Simple undersampling takes a random sample from the larger class, often the same size as the underrepresented class, in order to balance the classes. One concern with undersampling is that it doesn't take advantage of all of the data [151]. By definition, it discards a good portion of the over-represented class in order to provide class balance. It isn't clear undersampling is the best approach for highly imbalanced datasets, which may have anywhere from a 1:100 ratio to a 1:100,000 ratio for the minority class [170]. Agrawal [158] found no significant performance gain for undersampling across naive Bayes' classifiers, SVMs, and logistic regression models when predicting thirty-day readmissions for heart failure patients.

Oversampling

Oversampling randomly samples from the underrepresented class and adds this random sample to the same class. One issue with oversampling is overfitting the minority

class as it replicates minority class observations. Japkowicz [173] was able to show that oversampling can aid perceptron learning algorithms. Zolfaghar et al. [157] used oversampling in conjunction with logistic regression and random forests to predict all-cause unplanned thirty-day readmissions. The performance decreased when used with logistic regression and nominally increased performance with Random Forests but still underperformed the original logistic regression model. Agrawal [158] also performed oversampling and found no significant performance increase when predicting thirty-day readmissions over the original dataset for naive Bayes' classifiers, SVMs, and logistic regression.

Combinations of Under- and Oversampling Over- and under sampling can be combined in order to achieve class balance. The simplest approach is to achieve class balance by combining under- and oversampling directly. Meadem et al. [155] do exactly that to create their congestive heart failure readmissions models.

2.6.3 SMOTE

Many methods have been proposed to provide alternatives to under- and oversampling [174–179]. The most popular and well-used is the Synthetic Minority Oversampling Technique or SMOTE. SMOTE combines undersampling of the majority class with synthesized observations using linear regression techniques on a number of nearest neighbors [180]. Despite it being a well-established and widely used method to address imbalance, we have not found its use in the readmissions literature. Blagus and Lusa applied SMOTE to a high dimensional class-imbalanced dataset and compared its performance to simple under- and oversampling across many algorithms [181]. The algorithms include nearest neighbors, CART, LDA, QDA, Random Forests, support

vector machines, nearest shrunken centroids, and penalized logistic regression with both the linear and quadratic penalty. Their empirical results are based on simulated microarray data (1,000 variables). They found that SMOTE only improved performance for random forests without variable selection. Despite its improvement in performance, RF did not perform well overall. With variable selection, SMOTE was found to improve performance for the nearest neighbors and PAM algorithms but only performed well in the nearest neighbors cases for the high dimensional data. SMOTE was found to improve performance over the baseline for many of the low-dimensional cases.

2.6.4 Conclusions

Class imbalance issues have come to light in the machine learning research in the last fifteen years or so. The most popular methods for addressing class imbalance are simple under- and oversampling and SMOTE. Few heart failure readmissions studies have explored the benefits these techniques may offer. Those that have explored them have shown no notable improvement in predictive performance.

Chapter 3

Between Cohort Care Variation

This chapter presents a framework for working with administrative data for comparing all relevant procedures for a given diagnosis across cohorts defined according to the decisionmaking needs of the user. We present a methodology for using that framework and results from comparing select cohorts.

3.1 Background

Medical care variation research seeks to understand the differences in treatment patients receive, when those patients have the same diagnosis [182]. Care variation is an output of the diagnostic and treatment process occurring in the physician's mind. The inputs to the physician diagnostic and treatment process are many. Patient symptoms are not always obvious or correctly described by the patient or properly understood by the physician. Treatment options are many and are often applied in different order depending on the circumstances. Patient preferences influence the physician's choices. Finally, the physicians have varying abilities for interacting with patients for obtaining

symptom information, have differing abilities for observing and interpreting symptoms, and have different knowledge and beliefs (including biases [18, 61]) about appropriate treatment paths [182].

A typical care variation study selects a population based on some common feature (such as having the same diagnosis [31]) using a distinguishing feature by which to assess variation in some measure of treatment. Examples of distinguishing features include ethnicity [45], geography [25,35], types of physicians attending the patient [37], differing academic hospitals [30], or age group [21,36]. Treatment measures may include resources utilized to treat patients [183] or usage rates of certain medicines [24] or medical procedures [32]. Typically the study population is based on a diagnosis which has a treatment that is currently accepted to be the standard of care, such as the use of beta blockers and ACE inhibitors for many patients who experience congestive heart failure [33].

However, measures of variation across all relevant procedures for a given diagnosis are still lacking. We present a useful framework for working with hospital administrative data, a methodology for using that framework, and results from comparing some select cohorts.

3.2 Patient Care Model

We begin by discussing the basic process a patient goes through when receiving treatment from a care organization like a hospital. In the simplest form, a patient enters a hospital with some physiological conditions and then receives procedures to address those conditions. The care process is much more involved than this simple description and is approximated through a physician diagnostic process where physicians assign

diagnoses and patients receive corresponding procedures. This may be represented as a simple input-output model where the inputs are the patient diagnoses and the outputs are the procedures performed. Assessing the variation in this simple mapping from input (diagnosis) to output (procedure) is addressed through this research. Perturbations or variation in this mapping will be assumed to represent potential reductions or improvements in quality of care. Measuring these perturbations allows for further study and is a key step to understanding their true nature and thereby assuring better quality in patient care.

A More Complete Description The path from initial hospital entry to the creation of the patient electronic health record is not as straightforward as we initially suggest. Various inputs create the final record including the care provider's mental process; transcriptions of doctor-patient interactions recorded by scribes; clinicians, nurses, and others' recorded notes, diagnoses, procedures, and symptoms; trained coders who interpret the notes to assign ICD-9 and CPT codes for the actions performed on the patient (diagnoses and procedures); and others.

A patient enters the hospital (the black box system) with some unknown physical condition and underlying causes. The patient inside this black box is eventually assigned a set of diagnoses (sometimes symptom descriptions depending on the phase of treatment). These diagnoses feed back into the black box which then assigns procedures meant to address those diagnoses. A patient is assigned a principal diagnosis and a principal procedure for each visit. "The Principal diagnosis is defined as the condition established after study to have been chiefly responsible for occasioning the admission of the patient to the hospital for care . . . [and is] . . . represented by the ICD-9 diagnosis code [184]." The principal procedure does not have a definition of equivalent detail in the CDR and so is assumed to be the procedure which represents the prin-

cipal treatment given for that particular visit in response to the patient's condition. Secondary diagnoses and procedures are those which contributed to admission but are other than principal.

A Simple Input-Output Model of Patient Care It would stand to reason that the principal diagnosis and principal procedure are correlated. The actual recorded and reported principal diagnoses and procedures are assumed to represent in some respect the true description of physical condition and true response to that actual physical condition. It is also assumed that all patients are independent from each other.

We can represent the black box system as a basic input-output model with the diagnoses as the input and the procedures as output. It would be expected that in the ideal system identical inputs would result in the same outputs, or at least a very similar distribution of outputs. Assessing the variation in output (procedures performed) for given inputs (patient condition as reflected by diagnoses) is one of the contributions of this research.

Diagnosis-Response Matrix A data matrix or data frame can be used to represent this input-output relationship between diagnoses and procedures. The diagnoses may be considered the inputs (columns) and the procedures the response (rows) to the diagnoses. The values in the cells are some representation of the procedures like the number of times a given procedure was used in response to the diagnosis. A column then approximates the response to that diagnosis for the patient(s) in consideration. If the entries in the column are, say, zero for when a procedure is not used during a visit and one for when a procedure is used, then that column vector represents a response indicator vector. This vector could then be compared to other patients' vectors. Other metrics may be used as values to measure various phenomena of interest.

In this research these entries will be counts of how many times a procedure was used in response to the diagnosis. The matrix encodes the procedure responses to all the diagnoses for patients. The same diagnosis columns of the matrix from two separate groups of patients may then be used to compare the responses of the care system to each cohort. This matrix provides a way to represent and interpret the patient care process. We call this matrix the diagnosis-response matrix or DRM. All of the care variation analysis we perform is with the column vector of the DRM that corresponds to congestive heart failure, or ICD-9 code 428.0

Formal Description of the DRM This matrix is a $p \times d$ matrix where p is the number of unique procedures a population set has, and d is the number of unique principal diagnoses a population set has (See Figure 3.3 for a formal depiction and Figure 3.2 for a truncated example). The entries of the matrix are formed by counting the number of times each procedure (row) was indicated as the response to the given principal diagnosis (column). Other methods may be used to generate the entries to better answer research questions. Each diagnosis (column) has a distribution of procedures as a response to the given diagnosis.

The matrix may be created for an individual patient, in which case the matrix represents the response of the care provider organization to the patient's condition, or it may be aggregated by summing all matrices of patients in a specified sub-population for cross-population analyses. Each column of the aggregated DRM represents the response profile of the care provider organization to that population's conditions. Thus, for any given diagnosis the actual distribution of responses for that level of the organization are shown.

Further analysis could be performed using a similar matrix where the columns of the

$$DRM = A_{p,d} = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,d} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p,1} & a_{p,2} & \cdots & a_{p,d} \end{pmatrix}$$

Entries are procedure counts.

Figure 3.1: Formal depiction of the Diagnosis-Response Matrix. Note that the entries are procedure (row) counts in response to the corresponding diagnosis (column).

$$DRM = \begin{array}{c} \\ \textit{blank} \\ 01360 \\ 01402 \\ 33210 \\ 33225 \\ 36415 \\ 71010 \\ 71020 \\ 78465 \\ \vdots \\ 99309 \end{array} \begin{array}{c} \begin{array}{ccccc} 414.01 & 427.32 & 428.0 & \cdots & 996.67 \end{array} \\ \begin{pmatrix} 0 & 0 & 41 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ 0 & 0 & 2 & \cdots & 0 \\ 0 & 0 & 8 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 1 & \cdots & 0 \end{pmatrix} \end{array}$$

Figure 3.2: Example of the Diagnosis-Response Matrix. The codes over each column are ICD-9 codes indicating the diagnosis. The codes for each row are the procedures using Current Procedural Terminology (CPT). For example, the entry in column 3, ICD-9 code 428.0, and row 4, CPT code 33210 has entry 1 meaning that the procedure 33210, "Insertion or replacement of temporary transvenous single chamber cardiac electrode or pacemaker catheter" was performed one time in response to diagnosis 428.0, Congestive Heart Failure, in this patient group.

matrix become the procedures and the rows become the diagnoses generating a $d \times p$ matrix. This matrix is called the Procedure-Response Matrix (PRM). This matrix is interpreted differently from the DRM as it models how a procedure is used across various diagnoses. We do not use it in this research but mention it for the sake of possible future work.

$$PRM = B_{d,p} = \begin{pmatrix} b_{1,1} & b_{1,2} & \cdots & b_{1,p} \\ b_{2,1} & b_{2,2} & \cdots & b_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ b_{d,1} & b_{d,2} & \cdots & b_{d,p} \end{pmatrix}$$

Entries are diagnosis counts.

Figure 3.3: Formal depiction of the Procedure-Response Matrix. Note that the entries are diagnosis (row) counts in response to the corresponding procedure (column).

Using the DRM When thinking about variation in care using the DRM we look at variation in terms of all procedures responding to a single diagnosis. In terms of the DRM, this would be a column vector. Constructing the entire DRM is actually not necessary, a researcher need only construct the vector of procedure counts relevant to the diagnosis in question. The DRM is useful for characterizing the approximate input-output relationship of patients entering a care organization, receiving diagnoses, and then receiving corresponding procedures. Column vectors representing the same diagnosis from DRMs of patient cohorts are compared using an appropriate method to then measure the variation in the procedures used in response to the same diagnosis. This measure represents a measure of care quality in terms of variation between the two cohorts.

3.3 Methodology

The method compares two vectors of counts from two distinct patient cohorts. These cohorts could be patient groups with and without incidences of a certain comorbidity, procedures administered solely by one attending physician versus another, and so on. Table 3.1 shows an example of these vectors of counts that may then compared using the χ^2 test for independence discussed at length in Section 2.2.

	Diabetes II	No Diabetes II
Procedure 1	31	48
Procedure 2	11	18
⋮	⋮	⋮
Procedure P	56	57

Table 3.1: Example table of patient cohorts divided by having or not having Diabetes Type II. Each row is the number of times a procedure was used in the respective cohort.

The two-sample test of independence for measuring quality of care, in this case, care variation, has a nice interpretation. When comparing the diagnosis vectors from two cohorts the two-sample test of independence has the null hypothesis,

H_0, which is that the proportion of each procedure patients receive is independent of whether they have the stated comorbidity or not.

The alternative hypothesis is

H_a, which is that the proportion of each procedure patients receive is not independent of whether they have the stated comorbidity or not.

If we select our significance level α to be 0.05, then if

$p \leq 0.05$ we *reject* H_0, else if

$p > 0.05$ we *fail to reject* H_0.

Intuitively, if we reject H_0 we are saying that having or not having the comorbidity shows evidence of association with the proportion of each procedure patients receive. This means that there is evidence to suggest that the number of times the tested procedures are used may be associated with whether the comorbidity is present in the patients' histories or not. This is directly related to quality of care as this represents a measure of variation in the care, or number of procedures, patients receive. This significant difference in the proportion of procedures two cohorts receive may have clinical significance.

The next step in the methodology is to then take the difference in proportions each patient group receives relative to the other and rank them in order of highest to lowest. The extremes of this ranking represent those procedures which most differ in proportional application to each cohort. These ranked procedures are then presented to decisionmakers for further inquiry and determination of clinical significance. We perform these tests in `R` [106] using the `chisq.test()` function in the `stats` package which offers the option to simulate the p-value using the Monte Carlo method described in Section 2.2.2.

The overall process for measuring variation in treatment in response to a given diagnosis is:

1. Choose patient cohorts for comparison.
2. Choose diagnosis of interest.
3. Create diagnosis response vectors by counting the number of times a procedure is used across all patients in the cohort.
4. Compare the vectors using appropriate metric, E.g., χ^2 test for independence.
5. For each vector,

(a) Calculate the proportion each procedure is of the each vector.

(b) Take the difference of the two proportion vectors.

(c) Rank order the procedures in the differenced vector.

In the following sections, we present some cases where this methodology has been applied.

3.4 Results

This section presents results from several cases. The first case tests for independence between groups of patients with and without a given comorbidity in their patient history. This first case uses only the principal procedures. The second set of results differentiates between variation due to the principal procedure counts and that due to the secondary procedure counts. Within these first two cases we also provide evidence that less specific diagnoses may have a statistically significant relationship with greater variation between patient cohorts. The third section revisits some results found in the first two sections for atherosclerosis patients. The final set of results is a validation piece which tests for independence between patient cohorts being treated by two different physicians.

3.4.1 Between Diagnosis Group Care Variation Results Using Principal Procedures

We begin by observing the top ten most common comorbidities for heart failure patients in our population. Table 3.2 shows the top comorbidity is hypertension with

atrial fibrillation close behind. There are two versions of coronary atherosclerosis appearing in rank 3 and rank 8. The rank 3 diagnosis is more highly specified while the other is not. The fourth most common comorbidity is not a cardiovascular disorder; it is type II diabetes. The fifth most common comorbidity is an unspecified cardiomyapathy, or disease of the heart muscle. The sixth is a variant of the more specific heart failure diagnosis the population was constructed under (428.0), 428.9, which is another unspecified diagnosis. The seventh most common is unspecified hyperlipidemia. Ninth is pleural effusion which is fluid in the chest cavity. The tenth most common comorbidity is pulmonary congestion or fluid backup. Many of these are directly related to congestive heart failure with some that are not as direct, such as diabetes.

Table 3.2: The top 10 most common comorbidities for heart failure patients in our population ranked by number of times the comorbidity diagnoses appear.

Rank	Code	Description	Count
1	401.9	HYPERTENSION NOS	28888
2	427.31	ATRIAL FIBRILLATION	28537
3	414.01	CORONARY ATHERO NATIVE VESSEL	26736
4	250.00	DIABETES UNCOMPL TYPE II	20956
5	425.4	PRIM CARDIOMYOPATHY NEC	19227
6	428.9	HEART FAILURE NOS	18108
7	272.4	HYPERLIPIDEMIA NEC/NOS	17015
8	414.00	CORONARY ATHERO NOS	13137
9	511.9	PLEURAL EFFUSION NOS	12093
10	514	PULM CONGEST/HYPOSTASIS	11590

DRMs were created for each unique diagnosis using counts of principal procedures as a response to the given diagnosis. This means that each diagnosis had a DRM constructed from those patients who were diagnosed with the given diagnosis and another DRM was created from those patients who were not diagnosed with that diagnosis. This was performed across all 2383 diagnoses in the dataset. Each DRM contains columns corresponding to diagnoses and rows to principal procedures. Because this dataset is comprised of heart failure patients, ICD9 code 428.0 (Congestive Heart

Failure) is the main diagnosis of interest for testing variation in treatment.

Therefore, variation in treatment of congestive heart failure was tested using Pearson's χ^2 test. This means that in each DRM the χ^2 test was perfomed on the procedure distribution created by the column corresponding to diagnosis 428.0. These two distributions represent the distribution of procedures used to treat congestive heart failure for the respective cohort.

As we are testing for significance across a large set of cohorts, we correct p-values using the Holm-Bonferroni correction. We also record the number of times a comorbidity is significantly different across the division of patients with and without it. We perform the correction in an iterative manner. We begin by correcting the first two p-values together, then the first three together, then the first four, and so on until all 2,383 p-values have been corrected together. We keep track of how many times each cohorts pair has been found to be significantly different as this information may be useful for exploratory purposes. The second column in Table 3.3 shows how many times each diagnosis was significant out of the iterative Holm-Bonferroni corrections. This method proceeds through the comorbidities in order of commonality. Performing the Holm-Bonferroni in this order applies stricter criteria to later diagnoses adding an additional buffer against spurious findings. It could be argued that this is still too conservative [185].

Table 3.3 lists all diagnoses found significantly different at least one time, their original rank in terms of commonness, along with supporting information. Of the top ten most common comorbidities, three out of the ten are not significant: both coronary atherosclerosis comorbidities and type II diabetes. It is possible that the atherosclerosis cohorts are confounding any differences that may be found as patients from one cohort may be found in the other's without cohort.

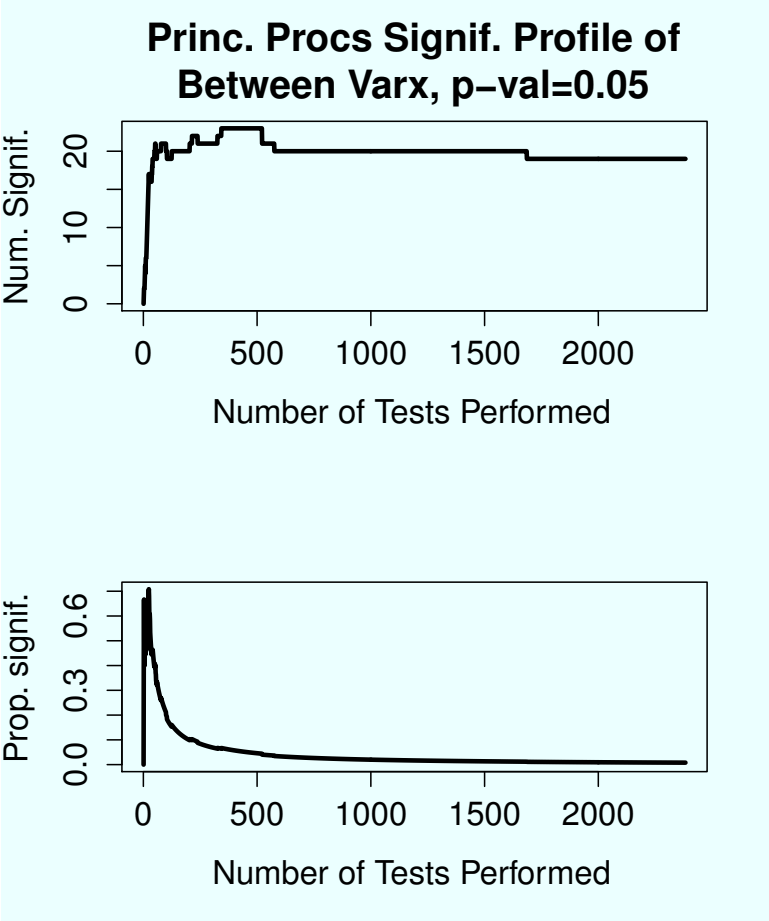

Figure 3.4: Significance profile as the number of cohort groups increases.

Figure 3.4 shows the results of the iterative Holm-Bonferroni corrections. The maximum number of corrected p-values that were significant with $\alpha = 0.05$ was 23 and it occurred first when 343 diagnoses' p-values had been corrected. This is shown in the upper part of the figure. The number of tests performed that resulted in the highest percentage of significantly different cohorts was after 24 tests with 70.8% of the 24 tests begin significant.

Table 3.4 shows that 19 diagnosis cohorts were still significantly different even after all 2,383 tests were corrected using Holm's procedure. Half of the top ten most common comorbidities still remain significantly different while hypertension, both athersclero-

sis comorbidities, type II diabetes, and the unspecified heart failure were not found significant.

Testing Significance of NEC/NOS Designation Some diagnoses indicated less certainty about the nature of the condition. These diagnoses are ones that may fall under a general classification but do not have a specific designation or subcategorization. These are denoted with the acronyms NEC or NOS which stand for 'Not Elsewhere Classified' and 'Not Otherwise Specified' respectively. We used these diagnoses to test whether there was greater variation in care between cohorts with more uncertain diagnoses and the corresponding cohort without those same diagnoses.

We counted the number of times a diagnosis contained one of these designations (NEC or NOS) out of the top 343 most common diagnoses. We chose the top 343 diagnoses as it was at 343 tests that the most diagnosis cohorts were found to be significantly different in procedure distributions using principal procedure counts. After 343 it stayed constant and then decreased slowly as more tests were run. Figure 3.4 shows the peak and slow decline of the number of significantly different cohorts. Using this number allowed the maximum number of potential diagnoses from either class (clear specification of diagnosis versus NOS/NEC).

Table 3.5 shows the contingency table obtained from the cohorts using only principal procedure counts. The number of diagnoses with the maximum number of significantly different cohorts was 343. We counted the number of times a diagnosis contained one of these designations (NEC or NOS) out of the top 343 diagnoses. Table 3.5 shows that there were 17 NOS/NEC diagnoses and only 14 that were more distinctly specified in the ICD-9 codes out of the significantly different group. Among those diagnoses that were not significantly different, 95 were NOS/NEC and 217 were not.

We used Pearson's chi-squared test for independence between the designation (NOS/NEC vs. specified) and the number of significantly different diagnoses. We obtained a χ^2 statistic of 7.6279, with 1 degree of freedom and a p-value of 0.006. This rejects the null hypothesis that the marginals are independent suggesting there is evidence of a relationship between finding diagnosis cohorts with principal procedures significantly different and whether they are unspecified diagnoses or not.

3.4.2 Between Diagnosis Group Care Variation Results Using All Procedures

It is possible that including all procedures designated as being used to treat 428.0 may affect the variation in care between cohorts of patients with and without given comorbidities. To test this we carried out the same analysis as in Section 3.4.1 but using the counts of all procedures rather than only the principal procedures. Table 3.6 shows those diagnoses that have significant variation between the diagnosis cohorts at least one time.

When all procedures are used to form the counts it increases the number of comorbidities with significantly different treatment procedure counts from 31 (when only principal procedures are used) to 79. Also notable is that all of the comorbidities up until the fourteenth are significantly different at least one time. Interestingly, the fourteenth *was* significantly different in the principal procedure comparisons.

Table 3.6: The number of times all cohorts were found significantly different when using all procedures labelled as addressing 428.0, what rank they were in how often they were used as well as their code and textual description.

Rank	Times Sig.	Code	Description	Count
1	2382	401.9	HYPERTENSION NOS	28888

Continued on next page

Table 3.6 – *Continued from previous page*

Rank	Times Sig.	Code	Description	Count
2	2381	427.31	ATRIAL FIBRILLATION	28537
3	2380	414.01	CRNRY ATHERO NATIVE VESSEL	26736
4	2379	250.00	DIABETES UNCOMPL TYPE II	20956
5	2378	425.4	PRIM CARDIOMYOPATHY NEC	19227
6	2377	428.9	HEART FAILURE NOS	18108
7	2376	272.4	HYPERLIPIDEMIA NEC/NOS	17015
8	2375	414.00	CORONARY ATHERO NOS	13137
9	2374	511.9	PLEURAL EFFUSION NOS	12093
10	2373	514	PULM CONGEST/HYPOSTASIS	11590
11	2372	496	CHR AIRWAY OBSTRUCT NEC	11063
12	2371	412	OLD MYOCARDIAL INFARCT	10318
13	2370	429.3	CARDIOMEGALY	10177
15	2368	427.1	PAROX VENTRIC TACHYCARD	8921
16	2367	786.05	SHORTNESS OF BREATH	8692
17	2366	518.0	PULMONARY COLLAPSE	8342
18	2365	793.1	ABN FINDINGS-LUNG FIELD	8004
19	2364	794.31	ABNORM ELECTROCARDIOGRAM	7502
20	2363	585.9	CHRONIC KIDNEY DIS NOS	7258
21	2362	786.09	RESPIRATORY ABNORM NEC	6887
22	2361	285.9	ANEMIA NOS	6529
23	2360	272.0	PURE HYPERCHOLESTEROLEM	6447
25	266	786.50	CHEST PAIN NOS	6399
26	2357	424.1	AORTIC VALVE DISORDER	6299
27	2356	424.0	MITRAL VALVE DISORDER	6266
28	2355	416.8	CHR PULMON HEART DIS NEC	5603
29	2354	428.22	CHRON SYSTOLIC HEART FAILURE	5443
30	2353	486	PNEUMONIA ORGANISM NOS	5200
32	2351	244.9	HYPOTHYROIDISM NOS	5117
33	2350	599.0	URIN TRACT INFECTION NOS	5110
35	5	518.81	RESPIRATORY FAILURE	4708
36	2347	530.81	ESOPHAGEAL REFLUX	4528
37	2346	427.9	CARDIAC DYSRHYTHMIA NOS	4516
38	2345	427.89	CARDIAC DYSRHYTHMIAS NEC	4423
42	2341	414.8	CHR ISCHEMIC HRT DIS NEC	4109
43	2340	327.23	OBSTRUCTIVE SLEEP APNEA	3825
44	2339	786.9	RESP SYS/CHEST SYMP NEC	3792
45	51	403.91	HYP RENAL NOS W REN FAIL	3546
46	2337	443.9	PERIPH VASCULAR DIS NOS	3540
47	2336	585.3	CHR KIDNEY DIS STAGE III	3407
49	2334	426.3	LEFT BB BLOCK NEC	3169

Continued on next page

Table 3.6 – *Continued from previous page*

Rank	Times Sig.	Code	Description	Count
52	150	311	DEPRESSIVE DISORDER NEC	3098
54	2329	518.89	OTHER LUNG DISEASE NEC	3027
56	2327	789.00	ABDOM PAIN NOS	2800
57	21	278.00	OBESITY UNSPECIFIED	2789
59	2324	593.9	RENAL & URETERAL DIS NOS	2366
60	2323	426.11	ATRIOVENT BLOCK-1ST DEGR	2256
61	2322	276.7	HYPERPOTASSEMIA	2232
64	496	250.02	DIABETES MELL TYPE II UNCONT	2028
65	2318	427.69	PREMATURE BEATS NEC	1956
66	2317	995.91	SYS INFLAM/INFEC W/O ORG DYS	1929
68	2315	585.4	CHR KIDNEY DIS STAGE IV	1911
69	23	276.1	HYPOSMOLALITY	1899
71	25	428.1	LEFT HEART FAILURE	1891
72	2311	276.8	HYPOPOTASSEMIA	1844
74	125	278.01	MORBID OBESITY	1810
76	537	426.4	RT BUNDLE BRANCH BLOCK	1758
77	2306	780.09	STUPOR	1748
82	2301	411.1	INTERMED CORONARY SYND	1622
86	2297	428.32	CHRONIC DIASTOLIC HRT FLR	1571
88	1218	578.9	GASTROINTEST HEMORR NOS	1528
89	184	571.5	CIRRHOSIS OF LIVER NOS	1521
90	267	287.5	THROMBOCYTOPENIA NOS	1519
93	30	790.7	BACTEREMIA NOS	1496
97	2286	276.51	DEHYDRATION	1419
99	2284	518.82	OTHER PULMONARY INSUFF	1395
116	241	008.45	CLOSTRIDIUM DIF	1216
121	27	586	RENAL FAILURE NOS	1123
126	53	428.23	ACUT ON CHRON SYSTOLIC HRT FLR	1064
138	475	786.59	CHEST PAIN NEC	938
141	1579	789.5	ASCITES	894
147	70	433.10	CAROTID ART OCCLUS W/O INFARCT	843
150	740	790.6	ABN BLOOD CHEMISTRY NEC	830
154	90	787.20	DYSPHAGIA NOS	807
163	352	437.0	CEREBRAL ATHEROSCLEROSIS	753
175	1545	793.0	ABN FINDING-SKULL & HEAD	654
197	2186	745.5	SECUNDUM ATRIAL SEPT DEF	576
233	245	E878.8	ABN REACT-SURG PROC NEC	463
342	2041	785.50	SHOCK NOS	292

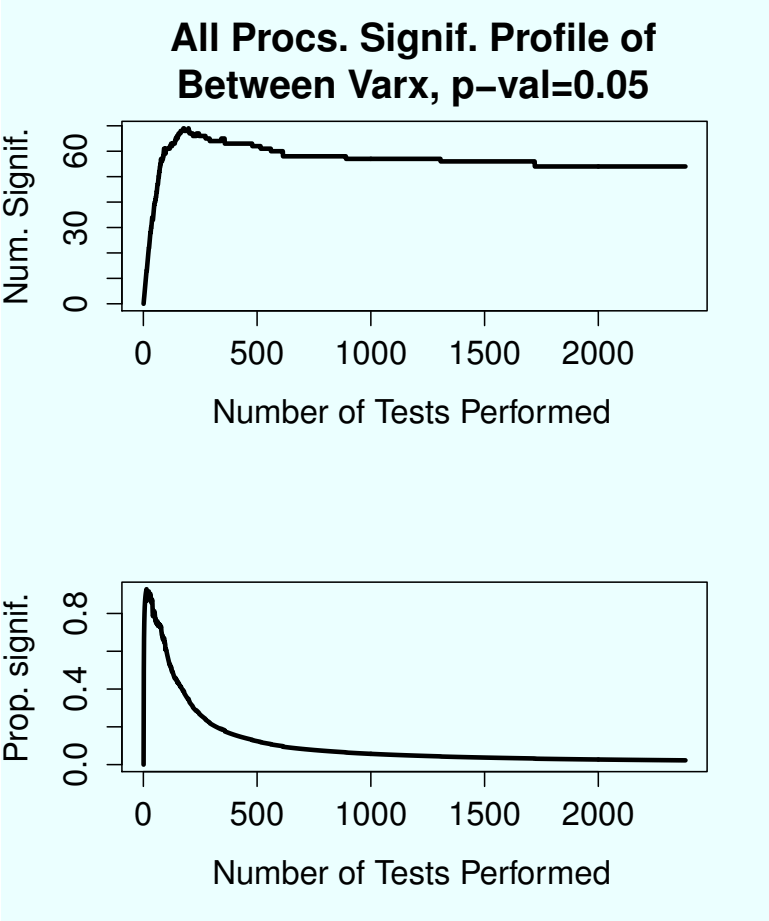

Figure 3.5: Significance profile as the number of cohort groups increases.

Figure 3.5 shows a similar pattern to that of the principal procedure's figure, where there is a sharp increase in both the number and percentage of significantly different procedure count distributions between diagnosis cohorts. The percentage drops quickly as fewer significantly different cohorts are found as less common diagnoses are tested. After all the tests are compared and the Holm-Bonferroni corrections are made, 54 diagnoses are still significantly different and are shown in Table 3.7.

Table 3.7 shows that 54 diagnosis cohorts were still significantly different even after all 2,383 tests were corrected using Holm's procedure. All of the top ten diagnoses remain significant throughout.

Table 3.7: Diagnoses found to be significantly different when even after Holm's correction for all 2,383 comparisons.

Rank	Code	Description	Count
1	401.9	HYPERTENSION NOS	28888
2	427.31	ATRIAL FIBRILLATION	28537
3	414.01	CORONARY ATHERO NATIVE VESSEL	26736
4	250.00	DIABETES UNCOMPL TYPE II	20956
5	425.4	PRIM CARDIOMYOPATHY NEC	19227
6	428.9	HEART FAILURE NOS	18108
7	272.4	HYPERLIPIDEMIA NEC/NOS	17015
8	414.00	CORONARY ATHERO NOS	13137
9	511.9	PLEURAL EFFUSION NOS	12093
10	514	PULM CONGEST/HYPOSTASIS	11590
11	496	CHR AIRWAY OBSTRUCT NEC	11063
12	412	OLD MYOCARDIAL INFARCT	10318
13	429.3	CARDIOMEGALY	10177
15	427.1	PAROX VENTRIC TACHYCARD	8921
16	786.05	SHORTNESS OF BREATH	8692
17	518.0	PULMONARY COLLAPSE	8342
18	793.1	ABN FINDINGS-LUNG FIELD	8004
19	794.31	ABNORM ELECTROCARDIOGRAM	7502
20	585.9	CHRONIC KIDNEY DIS NOS	7258
21	786.09	RESPIRATORY ABNORM NEC	6887
22	285.9	ANEMIA NOS	6529
23	272.0	PURE HYPERCHOLESTEROLEM	6447
26	424.1	AORTIC VALVE DISORDER	6299
27	424.0	MITRAL VALVE DISORDER	6266
28	416.8	CHR PULMON HEART DIS NEC	5603
29	428.22	CHRONIC SYSTOLIC HEART FAILURE	5443
30	486	PNEUMONIA ORGANISM NOS	5200
32	244.9	HYPOTHYROIDISM NOS	5117
33	599.0	URIN TRACT INFECTION NOS	5110
36	530.81	ESOPHAGEAL REFLUX	4528
37	427.9	CARDIAC DYSRHYTHMIA NOS	4516
38	427.89	CARDIAC DYSRHYTHMIAS NEC	4423
42	414.8	CHR ISCHEMIC HRT DIS NEC	4109
43	327.23	OBSTRUCTIVE SLEEP APNEA	3825
44	786.9	RESP SYS/CHEST SYMP NEC	3792
46	443.9	PERIPH VASCULAR DIS NOS	3540
47	585.3	CHR KIDNEY DIS STAGE III	3407
49	426.3	LEFT BB BLOCK NEC	3169

Continued on next page

Table 3.7 – *Continued from previous page*

Rank	Code	Description	Count
54	518.89	OTHER LUNG DISEASE NEC	3027
56	789.00	ABDOM PAIN NOS	2800
59	593.9	RENAL & URETERAL DIS NOS	2366
60	426.11	ATRIOVENT BLOCK-1ST DEGR	2256
61	276.7	HYPERPOTASSEMIA	2232
65	427.69	PREMATURE BEATS NEC	1956
66	995.91	SYS INFLAM / INFECTI W/O ORGAN DYSFUNC	1929
68	585.4	CHR KIDNEY DIS STAGE IV	1911
72	276.8	HYPOPOTASSEMIA	1844
77	780.09	STUPOR	1748
82	411.1	INTERMED CORONARY SYND	1622
86	428.32	CHRONIC DIASTOLIC HEART FAILURE	1571
97	276.51	DEHYDRATION	1419
99	518.82	OTHER PULMONARY INSUFF	1395
197	745.5	SECUNDUM ATRIAL SEPT DEF	576
342	785.50	SHOCK NOS	292

Table 3.8 shows those diagnoses that are significant when principal procedures are used but not when all procedures are used. The most common of the group is acute renal failure that is not otherwise specified which is the 14th most common comorbidity. The next most common is low blood oxygen or hypoxemia at rank 40. Edema is fluid buildup and is 48th most common with gout just behind in 51st. These may provide a good starting point for further exploration as to any patterns or clinically meaningful interpretations.

Table 3.9 shows those comorbidities that are significantly different when all procedures are used but not when only principal procedures are used. Of the top ten most common, both the atherosclerosis diagnoses are present as well as type II diabetes.

Table 3.9: Diagnoses found to be significantly different when all procedures are used but not found significantly different when only principal procedures are used.

Rank	Code	Description
3	414.01	CORONARY ATHERO NATIVE VESSEL
4	250.00	DIABETES UNCOMPL TYPE II
8	414.00	CORONARY ATHERO NOS
11	496	CHR AIRWAY OBSTRUCT NEC
12	412	OLD MYOCARDIAL INFARCT
25	786.50	CHEST PAIN NOS
26	424.1	AORTIC VALVE DISORDER
28	416.8	CHR PULMON HEART DIS NEC
29	428.22	CHRONIC SYSTOLIC HEART FAILURE
30	486	PNEUMONIA ORGANISM NOS
32	244.9	HYPOTHYROIDISM NOS
33	599.0	URIN TRACT INFECTION NOS
35	518.81	RESPIRATORY FAILURE
37	427.9	CARDIAC DYSRHYTHMIA NOS
42	414.8	CHR ISCHEMIC HRT DIS NEC
43	327.23	OBSTRUCTIVE SLEEP APNEA
44	786.9	RESP SYS/CHEST SYMP NEC
45	403.91	HYP RENAL NOS W REN FAIL
46	443.9	PERIPH VASCULAR DIS NOS
47	585.3	CHR KIDNEY DIS STAGE III
49	426.3	LEFT BB BLOCK NEC
52	311	DEPRESSIVE DISORDER NEC
54	518.89	OTHER LUNG DISEASE NEC
56	789.00	ABDOM PAIN NOS
57	278.00	OBESITY UNSPECIFIED
60	426.11	ATRIOVENT BLOCK-1ST DEGR
61	276.7	HYPERPOTASSEMIA
64	250.02	DIABETES MELL TYPE II UNCONT
65	427.69	PREMATURE BEATS NEC
66	995.91	SYS INFLAM / INFECTI W/O ORGAN DYSFUNC
68	585.4	CHR KIDNEY DIS STAGE IV
69	276.1	HYPOSMOLALITY
71	428.1	LEFT HEART FAILURE
72	276.8	HYPOPOTASSEMIA
74	278.01	MORBID OBESITY
76	426.4	RT BUNDLE BRANCH BLOCK
82	411.1	INTERMED CORONARY SYND
86	428.32	CHRONIC DIASTOLIC HEART FAILURE

Continued on next page

Table 3.9 – *Continued from previous page*

Rank	Code	Description
88	578.9	GASTROINTEST HEMORR NOS
89	571.5	CIRRHOSIS OF LIVER NOS
90	287.5	THROMBOCYTOPENIA NOS
93	790.7	BACTEREMIA NOS
97	276.51	DEHYDRATION
99	518.82	OTHER PULMONARY INSUFF
116	008.45	CLOSTRIDIUM DIF
121	586	RENAL FAILURE NOS
126	428.23	ACUTE ON CHRONIC SYSTOLIC HEART FAILR
138	786.59	CHEST PAIN NEC
141	789.5	ASCITES
147	433.10	CAROTID ART OCCLUS W/O INFARCT
150	790.6	ABN BLOOD CHEMISTRY NEC
154	787.20	DYSPHAGIA NOS
163	437.0	CEREBRAL ATHEROSCLEROSIS
175	793.0	ABN FINDING-SKULL & HEAD
197	745.5	SECUNDUM ATRIAL SEPT DEF
233	E878.8	ABN REACT-SURG PROC NEC

Table 3.10 shows those comorbidities that are significant in both procedure count methods, principal only and all procedure counts, are used to form the DRMs. Later entries in such a list may be indicative of comorbidities that are less thought about but potentially provide insight into treatment variation. Patients with shock seem to be treated significantly differently than those without shock. The clinical meaningfulness of this may provide helpful insight.

Each of these tables provides a different perspective on the comorbidities and how treatment differs between those patients with and without each of them. Principal procedures seem to be more discriminating with fewer overall diagnoses being significantly different. Despite the smaller number, some few are significantly different that are not significantly different when all procedures are used. When all procedures are used,

many more comorbidities differ in procedures counts between those with and without the comorbidities.

Testing Significance of NEC/NOS Designation with Counts of All Procedures As with the principal procedures, we found the number of procedures that yielded the highest possible number of significantly different cohorts. At 176 diagnoses being simultaneously tested, there were 79 significantly different diagnoses. The contingency table in Table 3.11 shows that there were 30 NOS/NEC diagnoses and 49 that were specified (as in assigned a specific ICD-9 code with a specific designation rather than NOS or NEC) out of the significantly different group. We then obtained the proportion of the top 176 diagnoses that were NOS/NEC versus those that weren't for the non-significantly different diagnoses.

Again, we use Pearson's chi-squared test for independence between the designation (NOS/NEC vs. specified) and significance for the contingency table constructed from the all procedures tests. We obtain a χ^2 statistic of 0.0045 with 1 degree of freedom and a p-value of 0.946. This fails to reject the null hypothesis that the marginals are independent suggesting there is not evidence of a relationship between finding diagnosis cohorts with all procedures significantly different and whether they are unspecified diagnoses or not.

This result is the opposite of that found when only principal procedures are used. This may imply that principal procedures reflect the treatment variation more sensitively when diagnoses are not adequately specified by either the ICD-9 (if a proper specification doesn't exist) or by the physician or coder (if it exists but isn't properly labeled).

3.4.3 Revisit: Variation Between Atherosclerosis Cohorts

We found in Section 3.4.1 that neither of the Atherosclerosis groups were significantly different from their corresponding cohorts without the same comorbidities. We posited that it could be due to the presence of the specified and non-specified atherosclerosis groups in each other's without cohorts. To test this we performed two tests. First we compared the two cohorts to each other using Pearson's Chi-squared test for independence with simulated p-value (based on 100,000 replicates). This resulted in a χ^2 value of 47.9 and p-value of essentially 1. This fails to reject the null hypothesis that the procedure counts are independent of the atherosclerosis groups. This suggests there is not evidence to suggest one comorbidity group applies procedures differently from the other.

Because they were not significantly different we combined them and compared the combined procedure distribution to that of the remaining patients. We tested them against the remaining patients using the same test yielding a χ^2 statistic of 240.7 and p-value < 0.001. This rejects the null hypothesis that procedure count distributions are independent of whether or not a patient is diagnosed with athersclerosis of a native vessel or not otherwise specified. This gives evidence to suggest that atherosclerosis patients experience different procedure application than patients without athersclerosis.

3.4.4 Validation: Between physician patient cohorts care variation results

Two physicians were chosen, here designated as physician A and physician B, because they performed the most procedures of all the physicians treating patients in

the population. Physician A performed 16,208 procedures while B performed 16,359 procedures. Physician A was the attending physician on 9451 visits while physician B was the attending on 5222 unique visits.

The patient cohorts were created by identifying those patient visits which were assigned one or the other but not both physicians as attending physician. To be sure the cohorts were comparable populations demographically, the distributions of Age, Race, and Gender were tested for independence using Pearson's χ^2 test. If the p-values are greater than $\alpha = 0.05$ then we fail to reject and the distributions are not statistically independent. This means that the groups are comparable and differences in treatment are not necessarily attributable to differences in demographics rather than the attending physician.

Pearson's Chi-squared test with simulated p-value (based on 2000 replicates) for testing similar distributions of Age across the two groups yielded $\chi^2 = 30$, and p-value $= 1$ meaning the groups are comparably comprised in terms of age distribution. Pearson's Chi-squared test with simulated p-value (based on 2000 replicates) for testing independence of Race distributions across the two cohorts yielded $\chi^2 = 28$ with p-value $= 0.1404$ meaning that the cohorts are also comparable ethnicly. Finally, the Pearson's Chi-squared test for independence of Gender distributions with simulated p-value (based on 2000 replicates) yielded $\chi^2 = 6$ with p-value $= 1$ meaning the cohorts are also comparable in gender composition. These findings allow us to state with more confidence that differences in treatment between the two groups are less likely to be attributable to demographic differences.

Pearson's χ^2 test with simulated p-value (based on 2000 replicates) across procedure distributions for physician A and B yielded results of $\chi^2 = 178.0206$ with associated p-value < 0.001 therefore rejecting the null hypothesis that the procedure frequency dis-

tributions are independent of the physician that performed them. This means that the procedure distributions between these physicians are significantly different and gives evidence to suggest that statistically significant variation in the amounts of procedures each physician uses to treat congestive heart failure exists.

To further explore the differences in variation, Table 3.12 shows the top five varying procedures performed by each physician in terms of percent difference. The top varying procedures for Physician A are major surgeries such as heart transplants. The top varying procedures for Physician B are minor surgeries, hemodialysis, and diagnostic in nature. These procedures may suggest a difference in training or specialty.

After performing these analyses we obtained the hospital service of the physicians in question and it turns out that Physician A specializes in cardiac surgery while Physician B does not specialize in cardiac surgery.

These findings validate our method. Ignorant of which types of physicians they were, and only knowing that they both performed a lot of procedures we found their patient cohorts were demographically similar while the procedures they used varied significantly. Identifying those procedures that each performed most (percentage-wise) more than the other it became clearer that each had a preference for choosing either intense surgery-based procedures, or less invasive procedures. Post-analysis we found that the physicians had different specialties both of which would naturally lead to the procedure choices reflected in our analysis.

3.5 Conclusion

3.5.1 Limitations

Some limitations of the method do apply. One assumption of χ^2 tests is that the cell counts are independent. This is not entirely true of procedures. Some procedures may be precursors to the next procedure performed having a serial correlation. While others may simply often be present in the performance of other procedures. These same statements may also be said of spoken language. This is an important comparison because independence of words is important in text analysis. This assumption is also required in many text mining methods such at latent semantic analysis [186] and latent dirichlet analysis [187]. However, in written word, the order words appear is *essential* to their making sense; this word order specification is called grammar. Nonetheless, it is common practice to make the "bag of words" assumption regardless of its gross violation in written language. In that same spirit, this method assumes the procedures are independent, and that physicians draw their procedures from a bag of procedures despite their being some sequential dependence (though arguably not nearly as strong as in language where it is formalized by grammar).

A second limitation is in the nature of the output of the χ^2 test itself. It does not highlight which cells contributed more or less to the rejection of or failed rejection of H_0. We address this by adding the ranking scheme which then allows users to explore which procedures may have clinical significance or other important meaning when the proportional application is different.

Table 3.3: The number of times all cohorts were found significantly different when using principal procedures labeled as addressing 428.0, what rank they were in how often they were used as well as their code and textual description.

	Times Signif.	Code	Description	Count
1	7	401.9	HYPERTENSION NOS	28888
2	2381	427.31	ATRIAL FIBRILLATION	28537
5	2378	425.4	PRIM CARDIOMYOPATHY NEC	19227
6	23	428.9	HEART FAILURE NOS	18108
7	2376	272.4	HYPERLIPIDEMIA NEC/NOS	17015
9	2374	511.9	PLEURAL EFFUSION NOS	12093
10	2373	514	PULM CONGEST/HYPOSTASIS	11590
13	2370	429.3	CARDIOMEGALY	10177
14	2369	584.9	ACUTE RENAL FAILURE NOS	9699
15	2368	427.1	PAROX VENTRIC TACHYCARD	8921
16	2367	786.05	SHORTNESS OF BREATH	8692
17	2366	518.0	PULMONARY COLLAPSE	8342
18	2365	793.1	ABN FINDINGS-LUNG FIELD	8004
19	2364	794.31	ABNORM ELECTROCARDIOGRAM	7502
20	5	585.9	CHRONIC KIDNEY DIS NOS	7258
21	34	786.09	RESPIRATORY ABNORM NEC	6887
22	2361	285.9	ANEMIA NOS	6529
23	498	272.0	PURE HYPERCHOLESTEROLEM	6447
27	2356	424.0	MITRAL VALVE DISORDER	6266
36	63	530.81	ESOPHAGEAL REFLUX	4528
38	1647	427.89	CARDIAC DYSRHYTHMIAS NEC	4423
40	13	799.02	HYPOXEMIA	4336
48	2335	782.3	EDEMA	3280
51	2332	274.9	GOUT NOS	3124
59	179	593.9	RENAL & URETERAL DIS NOS	2366
77	26	780.09	STUPOR	1748
124	2259	427.5	CARDIAC ARREST	1093
203	2180	780.96	GENERALIZED PAIN	545
213	362	070.70	HPT C W/O HEPAT COMA NOS	521
325	196	785.4	GANGRENE	321
342	2041	785.50	SHOCK NOS	292

Table 3.4: Diagnoses found to be significantly different even after Holm's correction for all 2,383 comparisons.

	Code	Description	Count
2	427.31	ATRIAL FIBRILLATION	28537
5	425.4	PRIM CARDIOMYOPATHY NEC	19227
7	272.4	HYPERLIPIDEMIA NEC/NOS	17015
9	511.9	PLEURAL EFFUSION NOS	12093
10	514	PULM CONGEST/HYPOSTASIS	11590
13	429.3	CARDIOMEGALY	10177
14	584.9	ACUTE RENAL FAILURE NOS	9699
15	427.1	PAROX VENTRIC TACHYCARD	8921
16	786.05	SHORTNESS OF BREATH	8692
17	518.0	PULMONARY COLLAPSE	8342
18	793.1	ABN FINDINGS-LUNG FIELD	8004
19	794.31	ABNORM ELECTROCARDIOGRAM	7502
22	285.9	ANEMIA NOS	6529
27	424.0	MITRAL VALVE DISORDER	6266
48	782.3	EDEMA	3280
51	274.9	GOUT NOS	3124
124	427.5	CARDIAC ARREST	1093
203	780.96	GENERALIZED PAIN	545
342	785.50	SHOCK NOS	292

Table 3.5: Contingency table comparing counts of NOS/NEC designated diagnoses that were found to be significantly different across cohort groups with the overall dataset counts in each designation. These results were obtained on the tests using only principal procedure counts.

	NOS/NEC	Specifically Designated
Significantly different diagnoses	17	14
Not significantly different diagnoses	95	217

Table 3.8: Diagnoses found to be significantly different when only principal procedures are used but not found significantly different when all procedures are used.

	Code	Description
14	584.9	ACUTE RENAL FAILURE NOS
40	799.02	HYPOXEMIA
48	782.3	EDEMA
51	274.9	GOUT NOS
124	427.5	CARDIAC ARREST
203	780.96	GENERALIZED PAIN
213	070.70	HPT C W/O HEPAT COMA NOS
325	785.4	GANGRENE

Table 3.10: Diagnoses found to be significantly different both when all procedures are used and when only principal procedures are used.

	Code	Description
1	401.9	HYPERTENSION NOS
2	427.31	ATRIAL FIBRILLATION
5	425.4	PRIM CARDIOMYOPATHY NEC
6	428.9	HEART FAILURE NOS
7	272.4	HYPERLIPIDEMIA NEC/NOS
9	511.9	PLEURAL EFFUSION NOS
10	514	PULM CONGEST/HYPOSTASIS
13	429.3	CARDIOMEGALY
15	427.1	PAROX VENTRIC TACHYCARD
16	786.05	SHORTNESS OF BREATH
17	518.0	PULMONARY COLLAPSE
18	793.1	ABN FINDINGS-LUNG FIELD
19	794.31	ABNORM ELECTROCARDIOGRAM
20	585.9	CHRONIC KIDNEY DIS NOS
21	786.09	RESPIRATORY ABNORM NEC
22	285.9	ANEMIA NOS
23	272.0	PURE HYPERCHOLESTEROLEM
27	424.0	MITRAL VALVE DISORDER
36	530.81	ESOPHAGEAL REFLUX
38	427.89	CARDIAC DYSRHYTHMIAS NEC
59	593.9	RENAL & URETERAL DIS NOS
77	780.09	STUPOR
342	785.50	SHOCK NOS

Table 3.11: Contingency table comparing counts of NOS/NEC designated diagnoses that were found to be significantly different across cohort groups with the overall dataset counts in each designation. These results were obtained on the tests using all procedure counts.

	NOS/NEC	Specifically Designated
Significantly different diagnoses	30	49
Not significantly different diagnoses	36	61

Physician A Procedures	%Δ	Physician B Procedures	%Δ
RT & LT HEART CATHETERS	9.4	DX ULTRASOUND-HEART	8.4
RIGHT HEART CATH	7.2	RT/LEFT HEART CARD CATH	6.8
RT HEART CARDIAC CATH	5.2	HEMODIALYSIS	6.3
HEART TRNSPLNT	5.1	THORACENTESIS	5.1
MPLNT CARD RESYNC DEFIB	1.9	EXC/DEST LES/TISS, OTH	3

Table 3.12: Percent difference in top 5 procedures administered by Physicians A and B

Chapter 4

Within Cohort Care Variation

While Chapter 3 addresses measuring variation in care between patient cohorts, this chapter assesses variation in care within a given cohort. Being able to measure the variation of care within a population provides a first step toward assessing how that variation impacts outcomes of interest that may reflect quality of care. This chapter presents measures of variation in application of procedures within a patient cohort. These measures are used in simple predictive models of visit charge. We report the methodology, results, and discuss.

4.1 Background

While many care variation research methods exist, very few methods provide a scalable approach for measuring variation in a way that facilitates insight into local treatment patterns. This highlights the two primary challenges preventing hospitals and other local decision makers from using medical and administrative "big data" for operational insights: *1*) local applicability of the data, *2*) scalability of the methods for both

generating and using the data for various levels of care organizations.

The first challenge derives from the nature of the data needed for typical studies. Data registries such as ADHERE [65], Optimize_HF [66], and SOLVD [67] house custom gathered data but aggregate over many hospitals and thus lose local applicability of their outcomes. Studies based on these registries focus on specific diagnostic groups and seek to measure variation in treatments, such as medications prescribed, but this quality of information is typically not available in the level of EHR system that 88% of US hospitals have [188]. This requires that data sources for such studies be custom created.

Locally applicable methods include qualitative studies that use chart reviews [54] and surveys or group sessions [55]– [60]. These locally applicable qualitative measures of care variation involve direct physician interaction for data collection, making scalability prohibitively time consuming.

To address these two challenges we define variation in a unique way, introducing the idea of visit complexity. Complexity is defined on the visit level while procedures have variation in application across patients. For example, if a procedure is applied at different rates for each patient in a population then that procedure has variation. If a visit is comprised of procedures which have high variation it is said to have high complexity.

Another way of saying that is that the procedure is applied uniformly. If a procedures' daily rate of application distribution (average number of times administered per day during a visit) is not uniform then it is said to have variation. Intuitively, procedures with higher variation are ones which have widely varying daily application rates. Measuring variation then must account for the magnitude in deviations in the dispersion of daily procedure application rates.

Our definition of variation relies on a couple of assumptions. The first is the assumption that there is some optimal sequence of procedures that should be applied to any patient given the same symptoms and comorbidities. This is the same assumption that underlies the formation of clinical practice guidelines, which prescribe standards of care for patients of certain diagnoses. In light of this assumption, variation is defined as applying either more or fewer procedures in a treatment sequence than that in the optimal sequence.

The second assumption is that a visit's complexity derives from a physician's search process. That is, the physician is still learning what the actual underlying physiological condition of the patient is, and before arriving at the correct diagnosis, administers intermediary procedures. These intermediary procedures may address symptoms or may be incorrect procedures appropriate for the hypothesized but ultimately incorrect diagnoses. Thus, a patient visit with greater complexity may have more procedures than the optimal visit. A visit may also have greater complexity if some procedures are neglected due to ignorance of some symptoms' causes or even presence. These two assumptions together motivate the hypothesis that complexity (and thus variation) correlates with outcomes of interest in a patient population.

The connection between measures of variation and pragmatic decisions on the local level hinges on whether procedure variation and visit complexity correlate with outcomes of interest. The research question we seek to address is about the nature of the procedures used in a visit: *do visits with procedures that on average have greater or less variation in application across a broadly defined population tend to correlate to visits' charges?* If the answer to the second question is yes, this validates the merits of the measures proposed in the first.

Affirmative answers to both questions allow us to utilized common billing data

for the purpose of understanding an organization's care variation patterns. Decision makers can then observe the patterns and focus on understanding those procedures or visit types that are most anomalous in terms of variation and complexity in the population of interest. While the proposed method does not explain *why* a procedure has greater or less variation, it does *measure* that variation in the population, allowing decision makers to prioritize which procedures or visits to further explore in order to understand what may be driving that variation. The goal of this paper is to validate a method that determines which procedures are most likely to incur increased hospital costs due to variation in their application, so that administrators can consider them more carefully. Key administrative changes such as policies, protocols, and practices can then be assessed over time using variation metrics and adjusted accordingly.

We propose that more information exists than that provided by the procedure counts alone. We believe that information is also captured in the visit complexity as derived from procedure variation across patients. To test this, we hypothesize that:

H_0 : Models including a measure for visit complexity have no difference in predictive performance for physician visit charge than a model with only procedure counts.

Rejection of H_0 suggests that our method can identify procedures that are more correlated with visit charge due to their inherent variation in application.

4.2 Data

The guidelines have nine major subdivisions of treatment by patients with these given comorbidities: hypertension with renal disease, hypertension without renal disease, myocardial infarction, anemia, diabetes, valvular disorders, coronary artery disease, atherosclerosis, and stroke. This paper reports analysis performed on the subset of all

patients who had at least one of the diagnosis codes for MI which include all codes under ICD-9 codes 410 and 412.

Table 4.1: Patient Subgroup Diagnosis Codes

Hypertension No Renal	401.xx, 402.xx, 405.xx
Myocardial Infarction	410.xx, 412.xx
Hypertenstion Yes Renal	403.xx, 404.xx
Anemia	280.xx, 281.xx, 282.xx, 283.xx, 284.xx
Diabetes	250.xx
Valvular Disorders	394.xx, 395.xx, 396.xx
Cornary Artery Disease	440.0x, 414.0x, 414.3x
Atherosclerosis	443.xx, 440.2x, 440.3x, 440.8x, 440.9x
Stroke	V17.1, 997.02, 434.xx

4.3 Methodology

Metric Calculation Process

We now describe how the raw data is transformed and used to measure procedure variation and subsequently visit complexity. For all patients and all procedures we create the standardized procedure count matrix, $S_{q,p}$ where $q = 1, \ldots, Q$ are the population's patients and $p = 1, \ldots, P$ are all the procedures present in the dataset, and each entry $s_{q,p}$ is the standardized count of each procedure across all of each patient's visits.

$$S_{q,p} = \begin{bmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,p} \\ s_{2,1} & s_{2,2} & \cdots & s_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{q,1} & s_{q,2} & \cdots & s_{q,p} \end{bmatrix}$$

Intuitively, each entry, $s_{q,p}$, is the average number of times the given procedure was administered per day for each visit it was administered for the given patient. Mathematically, $s_{q,p} = \frac{\sum_{v=1}^{V} p_v / l_v}{n}$, where $v = 1, \ldots, V$ is the set of all visits that contain

procedure p for patient, q, who in question; $q = 1, \ldots, Q$ is the set of all patients in the population; $p = 1, \ldots, P$ is the set of all procedures in the dataset; p_v is number of times procedure p was used in visit v; l_v is the length of stay (number of days) for visit v; and n is the number of patient q's visits containing procedure p.

Suppose a patient has three visits ($n = 3$) each comprising a single procedure which was administered different numbers of times each visit. Each visit lasts three, two, and four days respectively. Symbolically these are $l_1 = 3$, $l_2 = 2$, and $l_3 = 4$. The number of times the procedure was used in each visit was six times, five times, and ten times respectively or symbolically, $p_1 = 6$, $p_2 = 5$, and $p_3 = 10$. This means the average daily procedure usage rate, for this procedure for this patient was

$$\begin{aligned} s_{q,p} &= \frac{\frac{p_1}{l_1} + \frac{p_2}{l_2} + \frac{p_3}{l_3}}{n} \\ &= \frac{\frac{6}{3} + \frac{5}{2} + \frac{10}{4}}{3} \\ &\approx 2.334. \end{aligned}$$

So for the patient in consideration, that procedure was administered an average of 2.334 times per day during visits in which that procedure was used. The value, $s_{q,p}$, is calculated for each patient and each procedure filling out the standardized procedure count matrix, $S_{q,p}$, resulting in a matrix that would look something like the following:

$$S_{q,p} = \begin{bmatrix} 1.45 & .089 & \cdots & 0 \\ 2.2 & 0 & \cdots & .03 \\ \vdots & \vdots & \ddots & \vdots \\ 1.67 & 1.06 & \cdots & 0 \end{bmatrix}.$$

We then calculate the variation of each procedure (column of S) using three variation metrics: Standard Deviation (SD), Gini Index, and Normalized Entropy. These will be discussed in greater detail following the examples.

For example, suppose we have a dataset with three patients and for the first procedure their average daily application rates are in the column

$$S_{\cdot,1} = \begin{bmatrix} 1.23 \\ 3.9 \\ 2.7 \end{bmatrix}.$$

Then the Gini Index would yield a value of $Gini([1.23, 3.9, 2.7]) = 0.227$, normalized entropy would yield $NE([1.23, 3.9, 2.7]) = 0.085$, and $SD([1.23, 3.9, 2.7]) = 1.340$. These represent that procedure's variation based on the entire patient population in consideration (three patients in this trivial example). Completing these calculations for each procedure and all patients we now have a measure of procedure variation for each procedure in a patient group of our choosing.

We now return to the original dataset and view it by visit, where each patient may have multiple visits and each visit multiple days. For each patient visit we calculate the visit complexity by averaging the variation values for each procedure that comprises the visit. For example, suppose patient $q = 17$ has a visit with two procedures $p = 34$ and $p = 78$ which each have Gini index values of 0.23 and 0.78. The visit complexity in terms of the Gini index is then calculated to be $(0.23+0.78)/2 = 0.505$. We perform this calculation for each visit with each variation metric to obtain the study population's treatment complexities. This yields a vector of complexity values each corresponding to a visit.

As physicians are compensated based on the procedures they perform and record, the simplest and most obvious predictor of a visit's charge (without knowing anything

about the nature of the visit) is the number of procedures performed in a given visit. Intuitively, the greater the number of procedures performed in a visit, the greater one would expect the visit to cost. Therefore, the benchmark variable for comparison against the complexity measures is the simple number of procedures per visit. The final variable required for the method is the physician charge associated with each visit (or visit charge for short).

We are interested in understanding the variation in treatment, or visit complexity within the myocardial infarction subgroup of our chosen population which patients are, according the the CHFPG [189], supposed to be treated similarly. Therefore, we select those visits that correspond to patients who are members of the given comorbidity group. Membership is determined by doing a search based on the ICD-9 codes corresponding to those groups. If a patient has any visit with that code, they are a member of that subgroup.

In the original dataset each visit has a physician charge assigned. Each visit is also comprised of many procedures. To calculate a visit's complexity, we find the mean complexity of all procedures which comprise the visit. This results in a single overall complexity for the visit or $C_v = E[c_{p^i}]$ where $i = 1 \dots I$ are the procedures in a given visit. C_v is calculated for each visit in the dataset. This process yields the final dataset which comprises five variables: physician charge, number of procedures per visit, mean Gini coefficient per visit, mean standard deviation per visit, and mean normalized entropy per visit.

Table 4.2 shows the summaries of each variable. Visit charge is extremely right skewed. It turns out that the top quartile of visit charges accounts for ten times more in total charges than the lower three quartiles. The distribution of the number of procedures per visit is also right skewed.

Table 4.2: Variable summaries for final dataset.

	Visit Charge	Num. Procs. Visit	Norm. Entropy	Gini	SD
Min.	8.00	1.00	0.64	0.74	0.00
1st Qu.	332.00	1.00	0.99	0.77	0.05
Median	648.00	2.00	1.00	0.91	0.09
Mean	5462.00	6.44	0.99	0.88	0.12
3rd Qu.	2759.00	6.00	1.00	0.96	0.18
Max.	707000.00	272.00	1.00	1.00	0.23

Our variation metrics are the standard deviation, gini index, and normalized entropy of patients' average daily procedure application rates as described in Section 2.3. These metrics will form new variables by which we can predict outcomes of interest.

Models

To test our hypothesis we form 4 models:

1. M1: $\widehat{PhysicianCharge} \sim NumProcedures$
2. M2: $\widehat{PhysicianCharge} \sim NumProcedures + NormalizedEntropy$
3. M3: $\widehat{PhysicianCharge} \sim NumProcedures + Gini$
4. M4: $\widehat{PhysicianCharge} \sim NumProcedures + SD$

Each model regresses physician charge on the number of procedures with the last three models adding the visit complexity based on the respective variation measures to the model. Diagnostic plots show that normality and other assumptions necessary for least squares regression are not met even after transforming the variables. Therefore, we perform bootstrap robust linear regression.

We create 1,000 bootstrap samples and divide them into 2/3 training and 1/3 test sets. The training sets are used to fit coefficients. The resulting coefficients are applied

to the remaining test set to create physician charge predictions for each visit. The mean absolute error (MAE) is then calculated using the usual calculation. This yields four vectors of 1,000 MAE values corresponding to the four models. The models' results are shown in Table 4.3 which contains the mean bootstrap coefficients and accompanying 95% confidence intervals, LB and UB.

As expected, Table 4.3 shows that higher counts of procedures per visit correlate positively to a visit's charge. In the Myocardial Infarction subset, an increase in the number of procedures used in a visit increases the visit charge by about $1.17. When the normalized entropy based complexity measure is included in the model, its coefficient is strongly negative. As the normalized entropy is bounded [0,1] we interpret it slightly differently than the coefficient for the number of procedures used. An increase in 0.01 in entropy decreases (all else constant) visit charge by about $0.09 on average. The standard deviation based complexity metric also has a negative coefficient though of a much smaller magnitude. In contrast, Table 4.3 shows a positive coefficient for the Gini coefficient based complexity metric. A positive coefficient for the Gini metric implies that higher visit complexity correlates with higher visit charge. Negative coefficients for the Normalized Entropy and SD metrics imply that higher visit complexity correlates with lower visit charge on average. To interpret these we must remember to what parts of the distribution each metric gives weight. The Gini metric tends to weight higher values in the distribution less than the other two. The SD weights higher values more as it squares differences from the mean (which itself is also skewed by the higher values). Normalized entropy is also highly influenced by large outliers (and not nearly as influenced by small outliers).

In any case, the signs for each model coefficient indicate how to rank procedures for further study to reduce costs. Negative coefficients indicate that as visit complexity

Table 4.3: Model Results

M1: Benchmark	Mean Boot. Coef.	LB	UB
Intercept	5.866	5.829	5.905
Number Procs.	1.173	1.149	1.196
M2: Norm. Entropy			
Intercept	5.816	5.780	5.854
Number Procs.	1.151	1.126	1.175
Norm. Entropy	-8.622	-10.946	-6.398
M3: Gini			
Intercept	6.008	5.926	6.098
Number Procs.	1.133	1.101	1.164
Gini Coefficient	0.723	0.394	1.070
M4: Std. Dev.			
Intercept	5.464	5.349	5.565
Number Procs.	1.083	1.048	1.117
Standard Deviation	-0.206	-0.260	-0.151

increases visit charge tends to decrease. Positive coefficients indicate the opposite.

4.4 Results

The bootstrap coefficients' confidence intervals assumed a normal distribution of coefficients and the QQ-plot shown in Figure 4.1 verify this is true. Therefore, each of our coefficients' confidence intervals is valid and may be relied upon to determine statistical significance of the bootstrap coefficient values. It is possible, however, that if some of the coefficients' confidence intervals did include zero, the overall model could still outperform the base model. As the confidence intervals do not include zero we can conclude that the individual coefficients are statistically significant. To test model significance we review model predictive performance.

To test the null hypothesis, H_0, we compare the Mean Absolute Error (MAE) of each complexity model output to that of the simple model using a paired Wilcoxon

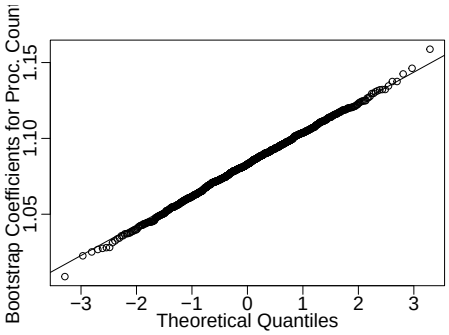

Figure 4.1: QQ-plot for the procedure count variable in Model 4. All four models' bootstrap coefficients had effectively identical QQ-plots.

signed-rank test. The paired test is appropriate because each of the four models is run on the same bootstrapped sample with observations in the same order before moving to the next sample in the sequence of 1,000 bootstrapped samples. As shown in Table 4.4, the normalized entropy reduced MAE the most at 0.029. The standard deviation reduced it by 0.004. However, the Gini coefficient did not outperform the base model as it had a mean increase in MAE of 0.002. Each of these results is statistically significant at $\alpha = 0.05$ as the p-values are less than 0.001 after Holm-Bonferroni correction [107].

Table 4.4: Mean differences in MAE between each complexity model and the base model. p-values from paired Wilcoxon signed-rank test with $\alpha = 0.05$.

Model	Mean Diff. in MAE.	p-value
Norm. Entropy	**-0.029**	< 0.001
Gini	0.002	< 0.001
SD	**-0.004**	< 0.001

When we use the MSE as our metric the practical significance becomes more notable. Table 4.5 shows that using both the normalized entropy and the standard deviation have a meaningful improvement in model prediction.

To verify the computational scalability of the method we tested how well the method

Table 4.5: Mean differences in MSE between each complexity model and base model. p-values from paired t-test where $\alpha = 0.05$.

Model	Mean Diff. in MSE.	p-value
Norm. Entropy	**-758**	< 0.001
Gini	66	< 0.001
SD	**-111**	< 0.001

performs in terms of runtime as the number of observations (visits) in the dataset increases. The method performance as measured by runtime (in seconds) is as expected with robust regression. Runtime increases linearly as the number of observations increase. We also tested the computing time on the MI subgroup changing the test set size and found that as test set size increases the runtime for the same sized group decreases approximately linearly.

4.5 Discussion

In two out of the three cases *we reject* H_0 using both MAE and MSE. For MI patients in this patient subgroup, the normalized entropy and SD complexity models outperform the base model. That these models outperformed the simple model gives evidence that they add information to procedures counts when predicting visit charges. This implies that measuring a visit's treatment complexity may provide insight into understanding how visit complexity is related to visit charges.

It is interesting that the metrics which improved predictive performance (normalized entropy and standard deviation) over the base model had negative coefficients. Negative coefficients for the complexity metrics implies that as complexity increases the visit cost decreases. As entropy tends to reflect outliers more strongly, its negative coefficient may reflect that outliers are not as costly as one may otherwise expect. And

as standard deviation also penalizes outliers, though to a lesser degree, its negative coefficient may imply the same thing. If the metrics do measure the parts of the visit charges distribution as conjectured, then this would also imply that visits with middle to high complexity tend to have lower costs than perhaps otherwise expected. This, however, needs to be confirmed by future work. Further research should be performed to more clearly determine the nature of the relationship between procedure variation, visit complexity, and visit charges. Such further research will also determine the extent to which these metrics may be used as practical tools. An example use would be to filter and prioritize procedures and/or visits for inspection and exploration to reduce costs.

We have shown that models predicting visit charge which include complexity measures outperform simple models which use only visit procedure counts. This implies that measuring visit complexity adds information to procedure counts for understanding what correlates with visit charges. The normalized entropy and standard deviation show a negative relationship with visit increase implying that as visit complexity increases, visit charge decreases. Future work is needed explore the properties of these metrics for decision making tools.

Chapter 5

Predicting Thirty Day Readmissions with No Clinical Data

This chapter is the first of several that explore variables and models for predicting unplanned thirty-day readmissions in heart failure patients. In this chapter we motivate why heart failure is of particular interest when addressing quality of care and explore models motivated by Chapter 4's results on visit charge using a data frame similar to the count matrix presented in Section 4.3. Random Forests models are presented using base variables, and models including the base plus procedure counts, diagnosis counts, and both procedure and diagnosis counts. These models are also adjusted using cost penalization for comparison. We present the methodology, results, discussion and conclusions. The majority of results in this chapter were published in [1].

5.1 Background

Heart failure is the most common reason for unplanned hospital readmissions. Rates of hospital readmission within thirty days are higher for congestive heart failure patients than for others [190] presenting a clear group of patients for which care quality could improve. Hernandez et al. [136] showed that following up with heart failure patients within seven days of discharge was correlated with lower readmissions within thirty days. While the ideal is to follow up with all patients, resource limitations do not allow it. It is necessary, then, to prioritize follow-up in a way that will minimize the number of readmissions. Accurately predicting those patients who are most likely to return within thirty days allows follow-up to be targeted most effectively. Such targeted follow-up could improve the quality of care of those patients which most greatly need an improvement.

However, typical thirty-day readmission prediction models either use data that are not readily available at the majority of US hospitals or use modeling techniques that do not provide adequate prediction accuracy (see Table 2.2 in Section 2.4.2). Moreover, the tendency of ongoing studies is to incorporate clinical data that are only present in the most modern electronic health record systems (EHRs). This is problematic as the population most affected by heart disease, the rural poor, is also the same population whose hospitals have the slowest adoption rates of advanced EHR systems.

To address these issues we apply the machine learning technique Random Forests to administrative claims data to predict unplanned all-cause thirty-day readmissions for congestive heart failure patients at UVA hospital.

5.2 Study Cohort

The query for creating the raw dataset for this research used the following general conditions: *a*) Principal/Secondary diagnosis: Congestive heart failure (ICD-9 code 428.0 [12]) *b*) Date of diagnosis: between January 1, 2006 and December 31, 2010. The data contained no patient identifiers and times were expressed as relative days rather than actual dates.

To define our study cohort we retained only those visits which were inpatient visits (19,189). We then removed all visits which had no readmission data recorded leaving 8,470 visits. For analytical traction we removed 219 visits that had patients of unknown gender, combined all ethnicities into one of three categories (other "O", black "B", or white "W"), and folded any payor with fewer entries than the "other" category into the "other" category. A summary of the final Payor distribution is in Table 5.1.

We tried to maintain as many unique Payor classes to allow for a greater number of classes for the analysis. Other studies aggregate payor to a much cruder level [131]. The tendency to aggregate possibly informative variables is the nature behind indices such as the Rothman Index [131], LACE and LACE+ scores [127], and Charlson score [191]. The practical purpose is for physicians to have a single number or score that provides meaningful information and allows them to make decisions quickly [131]. Our aim is to allow the algorithm to determine what is important and therefore we try to adhere to that principle as much as possible.

All ICD-9 supplementary "V" codes were removed along with all visits which had ICD-9 codes V72.8X (pre-operative exams) or V67.XX (followup visits) as a diagnosis. We then verified that all visits were unplanned according to the Centers for Medicare & Medicaid Services (CMS) criteria [124]. In short, each visit must not contain any of

thirty-two designated "planned" procedures. If they do have a planned procedure the visit may be deemed "unplanned" if one of twenty-six acute or complications of care conditions is present in the same visit. Our cohort met these criteria. A final removal of any visits or columns with no entries yielded the final cohort. The final cohort had 6,904 visits with 2,749 diagnoses assigned at least one time and 1,814 procedures performed at least once.

The outcome of interest was unplanned all-cause 30-day readmissions. The outcome was binary, either the patient returned in 30 days or less (value of 1) or they did not (value of 0). The age distributions shown in Table 5.1 show that the age distribution was not statistically significantly different (χ^2 test, p-value = 0.1417) between the two readmission groups. However, the mean age for the non-readmitted group was higher by 1 year (Wilcoxon rank sum test, p-value = 0.0143). The groups had no significant difference in gender (χ^2 test, p-value = 0.8721) or ethnicity (χ^2 test, p-value = 0.1858) distributions. The distribution of patients within the various payor classes did differ between the two groups (χ^2 test, p-value = 0.0017). Each Medicare and Medicaid program and the Blue Cross class had higher percentages of visits in the readmit group while the others had fewer. The length of stay distributions differed by a shift location of -4.305e-05 which while statistically significant (Wilcox rank sum test, p-value < 0.0001) isn't clinically meaningful. The medians for the two classes were the same (2) but the means differed greatly (21.8 vs 38.48).

20.9% of all visits were followed by a readmission within thirty days. This statistic is in line with those reported by the Dartmouth Atlas of Health Care for Congestive Heart Failure 30-day readmissions for the Charlottesville, VA, Hospital Referral Region (20.4%) and also across the entire Commonwealth of Virginia (20.9%) [9].

To test the predictive performance of hospital billing data we created four datasets.

Table 5.1: Cohort patient characteristics and relation to 30-Day Readmission (n = 6904)

		N (%))	Readmits N (%)	*p*-val.
Age	<45	580 (10.6)	160 (11.1)	0.1417
	45-64	1810 (33.1)	512 (35.5)	
	≥65	3073 (56.3)	769 (53.4)	
	Mean (SD)	65.4 (16.23)	64.2 (16.41)	*0.0143*
Sex	Female	2339 (42.8)	621 (43.1)	0.8721
	Male	3124 (57.2)	820 (56.9)	
Ethn.	Black	1538 (28.2)	441 (30.6)	0.1858
	White	3820 (69.9)	974 (67.6)	
	Other	105 (1.9)	26 (1.8)	
Payor	MCARE	3244 (59.4)	880 (61.1)	*0.0017*
	MCARE AD	270 (4.9)	74 (5.1)	
	MCAID	148 (2.7)	60 (4.2)	
	MCAIDHMO	223 (4.1)	48 (3.3)	
	BCROSS	310 (5.7)	96 (6.7)	
	SELF PAY	217 (4)	31 (2.2)	
	SO HLTH	130 (2.4)	29 (2)	
	OTHER	221 (4)	55 (3.8)	
	{BLANK}	700 (12.8)	168 (11.7)	
LOS	Med. / Mean (SD)	2 / 21.8 (37.57)	2 / 38.48 (45.05)	*<0.0001*
Total	Readmits ≤30d	5463 (79.1)	1441 (20.9)	

Each had as control variables Age, Gender, Ethnicity, Payor, and Length of Stay with either all procedures, all diagnoses, or both as additional variables. For example, a visit (observation or row) from the procedures dataset would have entries for each control variable and then the number of times each procedure was performed during that visit. A visit from the diagnosis dataset would have entries for the control variables and then the number of times each diagnosis was cited during the visit and recorded in the billing data. The both dataset had both the procedure counts and the diagnosis counts along with the control variables. The fourth dataset was the base dataset with only the control variables. The dataset was sparse as most visits had many fewer entries in

the variables than were possible.

5.3 Methodology

Our primary analytical method was Random Forests following many of Breiman's recommendations [159, 192]. A total of eight Random Forests models were generated. The four models were the base model, procedure model, diagnosis model, and both model. The base model consisted of only the control variables while the others were as described previously. The eight models are differentiated by the dataset used to form the Random Forests, and the prior weighting scheme on the response variable to compensate for the lack of balance. The models with a subscript p denote those using a prior weighting on the response variable and those without a subscript have no prior weighting. This results in the final eight models: *base*, *procedure*, *diagnosis*, *both*, $base_p$, $procedure_p$, $diagnosis_p$, and $both_p$.

We measured discriminative power of the models with the Area Under the Curve (AUC) (or c-statistic as the medical literature often refers to it). This metric is the most common used in the readmission literature both for congestive heart failure [124, 126, 130–132] and generally [148], therefore, for comparison's sake, we utilize it for our research as well.

Parameter Optimization

We began by randomly splitting the datasets into 2/3 training set and 1/3 test set. Each Random Forests model was built on the respective training set. Preliminary experiments showed that 500 trees yielded stable results.

Because the non-readmitted class of patient visits was so much larger than the readmitted class, we used prior weighting of the response variable, Readmit30, to more evenly balance the out-of-bag (OOB) and test error between the two classes. OOB error is the error between the response data values that is left out (thus out-of-bag) each time a tree is grown and the same data's fitted values. As Random Forests are designed to minimize overall predictive error, if one class is much larger than the other and the data are not very informative then the forests tend to perform better on the larger class and boost prediction error performance overall while sacrificing performance for the smaller class [159, 193].

For each classification model we began by setting a constant value of class weights where the 0s class had a weighting of 1 and the 1s class had a weight of 20. These values were held constant for each model. We then optimized mtry. At each step, a random sample of variables is selected as candidates to determine the next split. mtry is the number of variables randomly sampled as candidates for each split in a given tree. The noisier the variables, the larger mtry may need to be so that the likelihood of finding a good variable to partition the data increases [159].

To optimize the mtry parameter for the prior models we performed a two stage search. The first stage set mtry at Breiman's suggested values of $\sqrt{m}$ (default value), $.5 \cdot \sqrt{m}$, and $2 \cdot \sqrt{m}$ [159, 192], and then values from 100 to 1000 in increments of 100 resulting in thirteen initial test points. We created forests of 50 trees for each trial value. The errors typically became stable after only 20 or 25 trees. Each run provided several outputs. The first was OOB error for the entire dataset, the 0s class, and the 1s class. The second was the test set error for the same groupings.

Our objective was two-fold. First, was to find the value of mtry which best balances the OOB error between the entire set of responses, the 0s (non-readmits within 30

days) and the 1s (readmits within 30 days). The second objective was to see how well the OOB error represents the true prediction error as determined by the test set.

After finding a range of mtry values that yielded relatively stable OOB and test errors we repeated the search using 13 more mtry values within the stabler range of values. The final mtry parameter value was selected from this group based on reducing the test prediction error the most as well as evenly balancing the error between the entire response group, the 1s, and the 0s. The parameter search was different for the base model as it had only 5 variables and so we performed an exhaustive search for the optimal mtry value. As shown in Table 5.2, none of optimal mtry values for the prior weighted models were the default value. In fact, each optimal value was higher than the default.

Table 5.2: Stable ranges of mtry parameter for prior weighted Random Forests models.

Model	$\sqrt{m}$	Stable Range	Optimal mtry
$Base_p$	2	1-6	4
$Procedure_p$	43	220-279	279
$Diagnosis_p$	52	95-103	99
$Both_p$	68	117-137	130

For each tree, the Random Forests algorithm sampled from the training set with replacement and used out-of-bag (OOB) voting for classification of each visit. This yielded the OOB error estimate which we tracked and compared to the actual test error for each model. Table 5.3 shows that the OOB errors and test set prediction errors rarely differed by more than about two and a half percentage points. Additionally, in half the OOB error was greater than the test error and half it was less. The mean absolute difference in errors were all less than two percentage points.

Having established appropriate parameters for each dataset and model type we grew each forest and obtained variable importances. Our analysis was based on the OOB

Table 5.3: OOB and Test Set error comparison for total error, error for the not readmitted set (0s), and for the readmitted set (1s) error.

	OOB	0s	1s
$Base_p$	-0.59	0.58	2.51
$Procedure_p$	0.95	0.48	2.22
$Diagnosis_p$	-1.85	-2.15	-1.34
$Both_p$	-1.22	-1.88	1.31
Mean abs. diff in errors	1.15	1.27	1.84

importance as it is a measure of the decrease in accuracy as a result of a given variable's random assignment [159, 192]. We also calculated the gini importance and compared the rankings for the variables and found them to be similar. We opted for the decrease in OOB error as our importance metric as OOB error is an unbiased estimate of actual prediction error [159, 192].

Applying the models to their respective test sets, we obtained the test set AUCs and ROC curves. We also applied the models to 500 bootstrap samples of the original test sets in order to obtain an estimate of the AUC for each model with 95% confidence intervals.

5.4 Results

Table 5.4 shows that for each dataset the unweighted models have greatest overall performance in terms of bootstrap AUC. The non-prior models outperform the prior with a difference in mean bootstrap AUC of 0.08, 0.07, 0.06, and 0.04 for the base, procedure, diagnosis and both models respectively.

Table 5.4: AUCs for each model without and with prior weighting on the response and 95% confidence intervals.

	No Prior	95% CI	Prior	95% CI
Base	0.75	(0.72,0.77)	0.67	(0.64,0.69)
Procedure	0.75	(0.72,0.77)	0.68	(0.66,0.7)
Diagnosis	0.83	(0.81,0.85)	0.77	(0.74,0.79)
Both	0.84	(0.82,0.87)	0.8	(0.78,0.82)

5.4.1 Within No Prior Weighting

Interestingly, the base and procedure models have the same AUCs and confidence intervals. The ROC curves are essentially indistinguishable. The diagnosis model greatly outperforms the base with an AUC of 0.83. These differences are demonstrated with ROC curves in Figure 5.1. These ROC curves and associated AUCs were generated from the original test set. Of note is how long the false positive rate (1− sensitivity) stays low for each model as we increase sensitivity, especially the diagnosis and both models (begins to stray more quickly from the axis at about 0.4). The ROC curves for the base and procedure models reflect the bootstrap AUC results in that their performance is indistinguishable. The both model dominates all other models, though it dominates the diagnosis model by much less than the other two.

5.4.2 Within Prior Weighting

The procedure model outperforms the base model when more weight is given to the readmission class. A Wilcox rank sum test shows that the difference is significant (p-value <0.0001). The pattern continues with the diagnosis and both models greatly outperforming the remaining two. Notably, the both model outperforms the diagnosis model by a stronger margin (by 0.03) when prior weights to the response are applied.

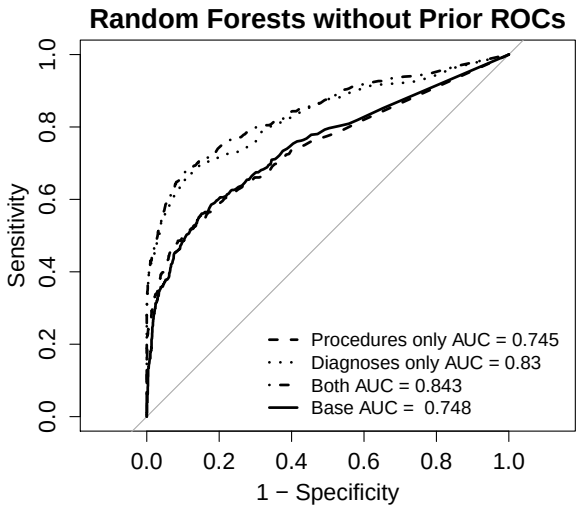

Figure 5.1: ROC curves and associated AUCs of each model using the default mtry value and without prior response weighting.

Figure 5.2 confirms that the weighting has the performance effect intended especially for the base and procedure models. The performance is near chance until about 0.25 on the false-positive axis where it seems to have an inflection point. Thereafter the models increase sharply and perform nearly as well as the diagnosis and both models. As designed, they sacrifice predictive performance on the 0s class for improved prediction of the 1s class. The decreased sensitivity at low false positive ranges is also noticeable for the diagnosis and both models.

5.4.3 Within Dataset Model Comparison

The predictive performance impact of weighting is more easily observed in Figure 5.3 particularly for the base and procedure models. For the base and procedure models the sacrifice of 0s performance in order to boost 1s predictive performance is indicated by the crossing of the prior curve past the non-prior curve at higher false positive rates. And from Figure 5.2 we see that it boosts their performance at about the same level

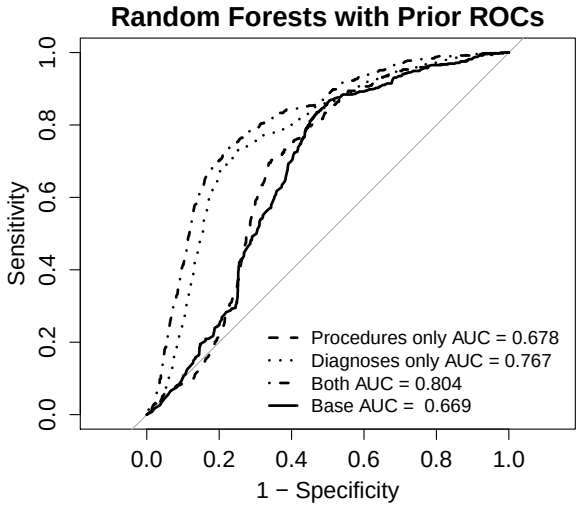

Figure 5.2: ROC curves and associated AUCs of each model using optimized mtry values and with prior response weighting.

as the other two models, within that range of sensitivity and specificity. The prior weighting does not markedly improve sensitivity over the entire range of sensitivities for the diagnosis and both models, rather, it reduces their performance in lower false positive ranges.

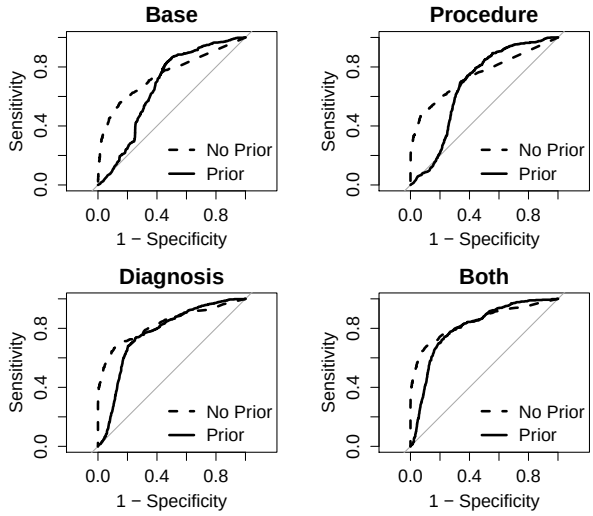

Figure 5.3: Comparison of prior weighted models to unweighted models.

5.4.4 Variable Importance

The top twenty ranked variables in terms of OOB error reduction importance are shown in Figure 5.4 (except the base models which have only 5 variables). For each non-prior model, the control variable length of stay (los) appears in at least the top eleven most important variables. It is most important for the base and procedure models and sixth in the both model. The interaction between length of stay and procedure variables was not explored but in models where both procedures and length of stay are present, it is higher. Payor also appears as the fourteenth most important variable in the non-prior procedure model and in rank fifteen in the non-prior both model. The same reasoning may apply to the payor variable as the length of stay: there may be some interaction with procedure variables. None of the other control variables were in the top twenty of any model.

In Table 5.5 we see that the both model's top 20 ranked variables were comprised of 7 of the procedure model's top 20 variables and 13 of the diagnosis model's top 20 variables. Length of stay was common among the procedure and the diagnosis models' top 20. The weighted procedure model shared 10 variables and the diagnosis model shared 8 variables with the both model. This sums to 37 unique variables which appear in the top 20 ranked variables for six different models.

The base model may reveal which control variables have influence on each outcome. Gender has the least influence in the unweighted model but highest on weighted implying that it may be more correlated to readmissions. Payor and age also swap order with payor getting less important in the weighted model. Length of stay appears to be important in each base model.

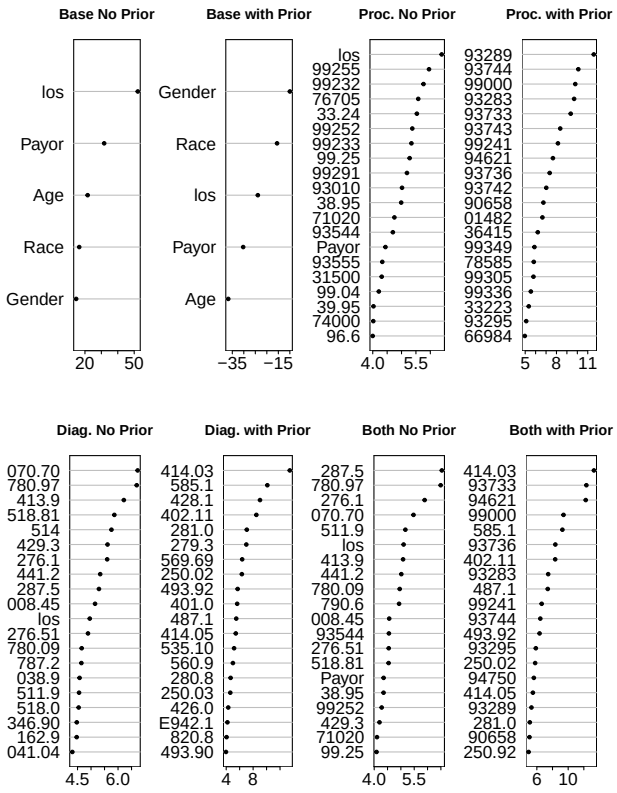

Figure 5.4: Twenty most important variables for the all models ranked by OOB importance.

Table 5.5: Overlapping variables for the both, diagnosis, and procedure models, weighted and unweighted.

Count	Proc-Both	Diag-Both	Proc-Both WT	Diag-Both WT
1	los	070.70	93289	414.03
2	99252	780.97	93744	585.1
3	99.25	413.9	99000	402.11
4	38.95	518.81	93283	281.0
5	71020	429.3	93733	250.02
6	93544	276.1	99241	493.92
7	Payor	441.2	94621	487.1
8		287.5	93736	414.05
9		008.45	90658	
10		los	93295	
11		276.51		
12		780.09		
13		511.9		

5.5 Conclusions

We found that with appropriate parameter selection, Random Forests can use simple billing data along with control variables to create highly predictive thirty-day readmission models. Across all tested datasets and weighting schemes, including the base model with only control variables, Random Forests may be used to form highly predictive models. Furthermore, these models may be used at virtually any hospital allowing those hospitals that are slower to adopt advanced electronic health record systems to more accurately predict heart failure thirty-day readmissions. With the best three models having an AUC of 0.8 or higher, our Random Forests models perform very well. Previous models, including those which use clinical variables not found in billing data, though not directly comparable, have not reported AUCs above 0.8 (see Table 2.2 in Section 2.4.2).

The best performing model of each type was the both model with the non-prior being the best overall. The both model included both diagnosis and procedure counts

as well as basic demographic data and payor. The jump in performance between the both models and the other models implies that interactions between the procedures and diagnoses may provide insight into factors leading to early readmissions. Interestingly, the addition of procedure variables did not improve performance in the unweighted case. However, when they interact with the diagnoses they give the both model a boost over the diagnosis model. It could be that procedure variables and length of stay convey the same information but when combined with diagnoses procedures provide more predictive information. The variables found in Table 5.5 show the top twenty variables identified by the Random Forests models as most important that are common between the various models. These variables may be used to further explore possible reasons behind early readmissions.

Though our study was not the first to utilize machine learning techniques for thirty-day readmission prediction, it was the first to use them for modeling with solely administrative data. Perhaps some of our performance gain over previous Random Forests studies was due to our optimizing the model parameters as no other cited studies reported optimizing model parameters. It could also be due to our variables being counts rather than just indicators. While we cannot verify that performance gain in the others' studies, we did verify it improved performance in our own.

The results in this study lend evidence to the importance of allowing the data to drive model and variable selection. We have shown that including all diagnosis and procedure variables, as well as demographic control data and length of stay, and allowing the machine learning method to select the variables to emphasize in the final models improves predictive performance over other models.

Our results are subject to some limitations. First, they were obtained on data from a single hospital and need to be verified at other locations. Additionally, coding of

procedures and data have been found to be inconsistent [194, 195] and these quality variations may lead to different results for other institutions. Additionally, we excluded all visits which contained no data for readmissions thereby capturing only a portion of the whole picture for our facility. Moreover, we did not model thirty-day mortality rates, which may be viewed as a form of readmission in the sense that patients were discharged under the assumption that health permitted it, yet they died within 30 days. While we compare the performance of our models with other models, because we have neither run their models on our data, nor our models on their data, the results are not completely comparable as generalizability is unknown.

Because our models have been built on widely available data, they may feasibly be used in helping hospitals without advanced EHR systems to prioritize discharge followup and thus reduce unplanned readmissions. This expands readmission research to those facilities which may benefit from it the most.

To conclude, we have confirmed that for our health system random forest analysis applied to ubiquitous billing and demographics data may be used to form highly predictive thirty-day readmission models.

Chapter 6

Thirty Day Readmissions Prediction Including Clinical Data

This chapter continues the work from Chapter 5 building Random Forests models on additional variables including clinical and non-clinical types. Two collections of Random Forests models are created. The first use the same control and derived variables as in Chapter 5 but add laboratory test results and medications administered during visits allowing for direct comparison in performance. The second collection of models takes these new models and adds two variables: an indicator variable of whether or not the visit considered is itself a thirty-day readmission and the number of prior inpatient visits. Performance results in the models are presented, compared, and discussed.

6.1 Background

Although many EHR systems do not have clinical variables readily available for analysis, their use continues to increase and it is important to assess their importance in predicting thirty-day readmissions. As more advanced EHR systems do come online, we should have proper models ready to improve those patients' quality of care. Properly prioritizing readmissions through prediction algorithms is a good example of care quality intervention that advanced machine learning can aid.

Interestingly, some have found that clinical variables do not improve such models [128, 129]. They do not go uncontested as other researchers have found that including clinical variables does improve predictive performance [127, 130, 131, 156]. We seek to address this issue by including clinical variables–specifically laboratory test results–and comparing models with and without those variables. In either case, it is important to know which models offer the most promise of improving patients' quality of care. We also include medications administered during inpatient visits as implicit clinical variables as they address patient symptoms. Models are created with labs and medications separately and then in combination.

6.2 Data

The data in this chapter uses the same data as in Chapter 5 with the addition of laboratory results and medications administered during patients' inpatient stay at UVA hospital. The data were created by obtaining all visits consisting of at least one diagnosis of 428.0, or Congestive Heart Failure, from a larger patient population where at some point in the patients' histories, they were diagnosed at least once with Congestive

Heart Failure.

Augmenting the billing data with lab results and medications administered provides data that may provide more information into the physiological state of the patient. Laboratory test results are explicit physiological measurements, including measurements of blood glucose, sodium concentration, and many others. Medications, however, are implicit metrics of a patient's physiological state. They are indicators of symptoms, or, to some degree, indicators of the actual condition. In both cases, the data provided may improve prediction of thirty-day readmissions.

6.2.1 Laboratory Tests

The laboratory results are prone to some biases. Biases mainly stem from the fact that the data are collected primarily for billing purposes. So non-billable actions are not accounted for. Some observations are missing results where a lab was performed but improperly recorded. The following lab tests had some missing entries and were removed: "CD4A" , "IGG" , "IGGC" , "IGGQ" , "KAPQ" , "SALBC" , "TESTOS" ,"UASN" , "UBAI" , "UGLN" , "UGLY" , "ULYS" , "USER" , "UTAU" , and "PTTC." An additional thirty-one of the lab results were marked as duplicates and were therefore removed.

Laboratory tests may be ordered multiple times per visit and visits may span multiple days. 37.5% (or 2587 out of 6904) of visits had the same labs run more than one time. 65.0% (4485/6904) of all visits had at least one lab test within the duration of the visit. Table 6.1 shows us a summary of the number of lab for any patient who had laboratory tests performed. One visit had 277 blood glucose level tests. Of all the labs run, 93.6% were runs beyond the initial request. This means that only about 6.4% of

all labs that were run (41881/653125 total labs run) were the original test.

Table 6.1: Summary of number of times labs were run.

	No. Times Run
Min.	1.00
1st Qu.	1.00
Median	1.50
Mean	9.58
3rd Qu.	11.00
Max.	277.00

In Table 6.2 we see what percent of the tests came out higher (H) than the normal range, lower (L) than the normal range, and within (N) the normal range. The majority, or 73.16% of tests came back indicating normal levels of the tested substance. If a test had a positive or negative result, 'N' indicates the negative result (as negative indicates a substance was in the normal range or not found) while 'P' indicates positive. Positive results account for the smallest percentage at 0.14% of all lab test results.

Table 6.2: Percent of labs that came out High, Low, or Normal.

	Test Result (%)
H	16.82
L	9.89
N	73.16
P	0.14

Table 6.3 show the most commonly run laboratory tests. The top three are albumin serum, blood urea nitrogen, and potassium. The ninth most common is glucose and is used for diabetes patients as well as for other monitoring.

6.2.2 Medications Administered During Visit

While the CDR [10] does not have prescription orders, it does have records of all the medications administered to patients during their inpatient visits. One major bias

Table 6.3: Top ten most commonly run laboratory tests.

Lab	Times Run
ALB	1966
BUN	1571
K	1445
CL	1443
CO2	1443
Na+	1443
CALCM	1441
CREAT	1441
GLUC	1440
GFRCAL	1354

is that the records were kept for billing purposes, so negative entries were used to indicate reimbursements [184], but have no medical meaning (you cannot receive a negative dosage of a pill, for example). Only 0.20% (or 993/490,773) had negative entries for the amount of doses of a given medication were administered. Additionally, 0.024% (or 117/490,773) of the administered medications had negative days indicating administration before actual admission. We have omitted them for consistency of interpretation. The final medication dataset had 490,773 entries with 2,295 different entries for medications.

Another source of bias is multiple entries representing different dosages of the same medicine. For example, Acetaminophen comes in multiple dosages such as 300mg or 600mg. Many medications had the same issue, as well as combinations with other drugs. For simplicity, the medications were left as is, so multiple entries exist for treating a patient with Acetaminophen. An argument may be made that the dosages contain different information, as using a single large dose of a pain killer may provide different information than multiple smaller dosages which all sum up to the same overall dosage in a given time frame.

Table 6.4 shows that the two most commonly used medication are both gastroin-

testinal. The most common, Docusate Sodium, is an anti-constipation treatment and the second most common, Esomeprazole, is used to address acid reflux. Of note is the fourth most common which is insulin for diabetes patients. The sixth most prevalent, Fentanyl, is a powerful pain killer while the seventh, Heparin, is a common blood thinner. Simvastatin is used to treat high cholesterol. The tenth most common, Furosemide, is used to treat edema, or excessive swelling.

Table 6.4: Top ten most commonly used medications.

Medication	Count
DOCUSATE NA CAP 100MG UD	11592
ESOMEPRAZOLE CAP 40MG UD	10742
ASPIRIN BABY 81M	9642
INSULIN REGULAR DOSE (A)	8616
SODIUM CHLOR .9% 1L BAG	8422
FENTANYL 2ML	7580
HEPARIN VL 5000 U	7529
SIMVASTATIN TAB	6445
ASPIRIN TABLET 3	6362
FUROSEMIDE TABLE	6343

6.2.3 Variable Creation

The analysis performed relies on several derived measures based on the original data. Part of our intent was to create variables that may yield insight into thirty-day readmissions as well as improve the models' predictability.

Derived Lab Variables

Because laboratory tests can be run multiple times during a visit some visits had more than a single result that could be reported as the visit's result. This required an aggregation of the lab results that would still retain as much information as possible

while allowing for a single number to summarize the lab results for that test during that visit. Lab results typically have either a normal range or a positive/negative result. For many lab tests, the lowest normal range value possible is zero, meaning there is no value that is below the normal range, but can only be too high. Other tests can be too low but not too high. And finally others can be both too high and too low. We developed several candidate metrics to try and capture the information available from each of these scenarios.

Minimum Lab Result Some labs have results that imply deteriorating physiological condition with lower than normal values reported. To capture this phenomenon we retained that lab result which was the minimum reported result over the course of the visit. This could be seen as a one sided test that only accounts for those tests which are worrisome only when the value is low. However, it could be the case that some patients' lowest results are higher than the normal range and still be reported.

Maximum Lab Result Some lab results indicate declining health when their test values are above the normal range. These instances were captured by a metric which retained only the maximum value of the reported lab results over a visit. This may also be seen as one sided but as conjectured before, a visit's maximum lab may still be in the lower range, which would still indicate declining health as measured by that test.

Indicator of Result Another metric is the categorical variable indicating the result, whether it was High (H), Low (L), Normal/Negative (N) or Positive (P). This variable indicates if the result came above the normal range (H), below the normal range (L), within the normal range (N), negative–meaning it did not indicate an undesirable result

and is thus normal (N), or if the lab test did indicate an undesirable result (P).

Final Visit Result Because the decision to discharge is often informed by the lab results reported just before a patient is thought well enough to be discharged, the visit's final lab result was also reported as a variable. This variable is simply the last value for each test. If a test was only run once, then that value is reported in this variable.

Number of Times Lab Run A count of the number of times a lab was run could indicate a few things. It could demonstrate that a certain test was important in updating the physician on the condition of the patient. It could also represent a patient with a chronic illness like diabetes who regularly has blood work done. The overall idea captured is that more lab runs may indicate more thought and concern for the well-being of the patient than patients with fewer lab tests run.

Sum of Squared Abnormality The intent of this final derived lab metric was to capture the full deviation of the lab results from the normal range across the entirety of the visit. Inspired by the typical loss function for linear regression, sum of squared residuals, this metric squares the difference of each result from the normal range and sums across all of them. This means that a result falling within the normal range will have an abnormality of zero and thus contributes nothing to the sum of squared abnormalities. But as values fall farther out of the normal range they are penalized more strongly due to the squaring.

An example may clarify. Suppose the normal range for lab test ABC is 5 to 10 mg/L and our patient has the test run 3 times with the results being 3, 15, and 7. First we calculate the abnormality of each result which is $(\min(N_l) - R_l, R_l - \max(N_l))_+$,

where N_l is the normal range for lab test $l$and R_l is the lab result for lab test l. In the case where no entry is positive the result is equal to zero. So the abnormality of each would be $(5-3=2, 3-10=-7)_+ = 2$, $(5-15=-10, 15-10=5)_+ = 5$, and $(5-7=-2, 7-10=-3)_+ = 0$. So the final abnormalities are [2, 5, 0].

These abnormalities are then squared and summed to give the squared abnormality of the lab across the entire visit. Formally this is $\sum_{i=1}^{I} [(\min(N_{li}) - R_{li}, R_{li} - \max(N_{li}))_+]^2$, where $i = 1, \ldots, I$ are the indices for each run of lab l, and $l = 1 \ldots L$ are the indices for the L potential labs in the dataset.

This variable is intended to capture the magnitude or severity of the abnormality of the lab results over the course of the visit. This may then reflect the severity of the physiological condition of the patient as measured by the given labs.

Derived Inpatient Medication Variables

Patients often receive medication or fluid during their inpatient stay for stabilization and symptom treatment. The dosages may vary or the medications adjusted as the patient's condition changes and this may be reflected in the amount or type of medication administered. Medications are coded according to dosage in the CDR, so a patient may receive an initial dosage of Acetaminophen of 600mg and then thereafter 300mg every few hours. This would be recorded as two separate medications, where each administration would be counted. So a patient with this administration would have '1' under ACETAMINOPHEN 600mg, and if they only received the 300mg dose twice during a given day the entry under ACETAMINOPHEN 300mg would have '2'. So the entries for medications are counts, or the number of times a stated dosage and medication/fluid were administered for each day throughout a patient's visit. This means that a patient potentially has multiple entries per visit per medication, corre-

sponding to each day the stated medication and dosage were administered. Similar to the laboratory test results, we created derived variables to capture the information inherent in these counts across days in a visit.

Maximum Administration Our first variable was the maximum number of times in any day a given medication was administered. This variable was thought to capture the implicit maximum severity of the symptoms the medication was intended to treat.

Number of Days Administered The next two variables capture similar but subtly distinct types of information. The first is the number of days during the stay which the medication was administered. This gives some idea of whether the medication was treating something acute or temporary, or something that may be ongoing. For example, if a patient's length of stay was 10 days but a medication was only administered 7 of those days, then perhaps that medication was for a temporary symptom, or was intended to stabilize the patient rather than be an ongoing therapy.

Number of Medication Administrations The second is the number of times a medication was administered throughout the visit. So a medication could have been administered a total of 4 times for 7 days out of a 10 day visit resulting in a variable value of $4 \times 7 = 28$. This would differ from the number of days administered above which would only be 7.

Standard Deviation of Administrations This is the first of three metrics used to capture the variation in the amount of daily administration. Variation metrics of medication administration may yield insight into the instability of the patients condition. This may be most strongly revealed in longer visits where the patient may

stabilize after a few days then suddenly decline and need a drastic increase of certain medications. Swings in medication changes would be revealed by such metrics.

Entropy of Administrations Entropy is a measure of the uniformity of a vector of numbers. This is a second candidate metric used to measure how uniformly medications are applied across days in the visit. The actual metric is normalized entropy as defined in Section 2.3.

Gini of Administrations The final derived medication variable is the gini measure across daily administrations of medications. This metric is widely used to measure inequality in wealth distribution and here measure inequality in the daily distribution of medications. This is a final candidate for measuring the variation/uniformity of medication administration throughout a visit.

Aggregated Metrics

The derived laboratory and medication variables described above create a large number of variables for any machine learning algorithm to account for. In addition, these data frames are sparse, as a given visit only uses a small fraction of the thousands of medications and laboratory tests available. To compare low-dimensional, information-dense data to sparse, high-dimensional datasets, we create aggregated metrics based on some of the derived variables. These metrics are usually a count aggregated from the derived metrics. We also created aggregate variables from the procedures and diagnoses.

The first two variables differ from the final six in that they are not derived from variables that refer to medical action (like administering a procedure, declaring a diag-

nosis, or giving a medication). They are variables which describe or aggregate no part of the patient's condition, rather state a fact about the nature of the current visit and prior visits. These variables were created after [1] was published and are not included in the first collection of models reported here to allow for direct comparison to the results in [1]. The number of procedures and number of diagnoses are the aggregated metrics for the data used in the original paper [1] and were also not in that paper. This is the first time they are used for analysis. The intent of the aggregated variables is to compare create models with simple aggregated metrics and compare their performance to models with many more variables and greater complexity. The results of these comparisons may reveal something about the nature of what we are measuring to predict unplanned thirty-day readmissions in heart failure patients. Simple models could be implemented easily to improve quality of care where computing resources are very limited. The results may also reveal the ability of various algorithms to handle different variable types and reveal what types of variables are best for predicting unplanned thirty-day readmissions.

Number of Prior Inpatient Visits This variable is the count of previous inpatient visits to the hospital that are on record and in the dataset. This means that for some, the number of prior visits will be truncated as those patients' records extend before the dataset was created.

Thirty-Day Readmission Indicator A novel variable created for this analysis was a binary variable indicating whether or not the present visit was itself an unplanned thirty-day readmission.

Number of Procedures This variable is a simple count of the number of procedures a patient receives during a visit. If a procedure is performed multiple times it is counted each time.

Number of Diagnoses This variable is a simple count of the number of diagnoses (principal and otherwise) a patient received in the duration of their visit.

Sum of Number of Procedures and Diagnoses This variable sums the prior two in an attempt to capture in one number how much activity was happening with regards to the patient's care.

Number of Labs Performed Perhaps the simplest and least explicit of the lab metrics is the count of how many labs were performed over the visit. As opposed to the other metrics, this results in a single number for each visit, whereas the other metrics result in a vector of numbers with each lab test having a number (most often zeros as most of the thousands of labs are not run in a given visit). This metric could be thought to capture the uncertainty in the patients physiological condition, or even the importance of monitoring. If many labs are run, that may indicate the importance of knowing the physiological state across time as it may greatly affect the patient's recovery. It could also indicate illnesses that need regular blood work done anyhow, such as diabetes mellitus, which is known to complicate many other conditions. This metric is calculated by counting the number of laboratory tests ordered during the visit.

Total Number of Medicated Days This variable is the total number of days all medications were administered. This is the sum of the number of days each medication

was administered. For example, if a patient was given Ibuprofen 200mg for 2 days and Acetaminophen 300mg for 5 days across a 10 day visit the total number of medicated days would be $2 + 5 = 7$. This aggregated metric is a reflection of how medicated the patient was during the stay.

Sum of the Sum of Squared Abnormalities This visit sums up the individual sums of square abnormalities across all of the lab results in a patient's visit.

6.3 Methodology

To train our models we randomly divide our data set into two-thirds training set and one-third test set. We do this 50 times. These same 50 training and test sets are used to compare all models and allow for repeatable, direct pairwise comparison. For each training set we use the default number of variables sampled at each node as it has been shown to offer the best performance on this kind of data [1]. We create 500 trees for each model which had also shown stable results.

As discussed in Section 2.4.2 the metric used for model selection is area under the curve or AUC. To select the final subset of variables we run Random Forests twice for each candidate model. We ran candidate models for each derived variable set with a base variable set found to perform best in [1] which included control variables, counts of procedures and counts of diagnoses.

Each model was run 50 times as described and the average variable importance was calculated. The variables were then sorted and those variables not greater than the absolute value of the lowest variable importance (generally a negative number) were dropped. This was done to quickly remove noise variables. More stages could be added.

Because of computational time and the sheer number of analyses being performed only one such cut was used. Also, completing another round of variable elimination did not improve the AUCs and so only one round of variable elimination was completed for all models. The model was then run another 50 times with only the remaining variables. The variable importances of the second batch of 50 were then averaged to obtain a final importance score and ranking of each. After each of the derived models was run, we also ran a model where all variables were aggregates of the derived variables. For this model, only one run was performed as all variables had high importance. To test whether the AUCs for one type of model were statistically significantly different from another we implemented the one-way (>) Wilcoxon signed rank test [196] using the Holm-Bonferroni correction for multiple comparisons [107].

6.4 Results

6.4.1 Modified Billing Variables

Prior to testing our hypotheses regarding clinical variables, we test a few variants of the billing data to show that diagnosis and procedure counts provided the largest improvement in performance. To do so we used the "both" dataset from Chapter 5 where the procedure and diagnosis counts were replaced by an indicator variable, or each of the variation metrics described in Section 4.3. This resulted in four models.

The hypothesis was that the count metrics used in Chapter 5 did not significantly improve model performance over the variation metrics or indicator variables. Another hypothesis is that none of the four variable types was significantly better than the others in terms of AUC. Rejecting these hypotheses supports the use of the counts as

proposed previously moving forward.

The Holm-Bonferroni adjusted p-values for testing the first hypothesis shown in Table 6.5 show that in each case the count data (AUC = 0.8417) used in Chapter 5 performs significantly better at an $\alpha = 0.001$. The second best model was the standard deviation with an AUC of 0.8384. The SD model had a Holm-Bonferroni adjusted p-value of 0.0589 when tested against the indicator data. The SD variable also outperforms the indicator variable by 0.0005 with p-value of 0.059 which is significant when $\alpha = 0.1$. The Count model's mean increase in AUC across the other four was 0.0041. After establishing that the count variables are best for billing data alone. We proceed to compare those models with models including clinical variables.

Table 6.5: Average AUC for first and second runs of the "both" data using variation metrics and indicators compared with the count variable from [1]. The p-values are from a one-sided Wilcoxon rank sum test against the Count AUCs.

Model	Run 1	Run 2	p-value
Indicator Variables	0.8141	0.8379	< 0.001
Entropy Variation	0.8141	0.8369	< 0.001
Gini Variation	0.8137	0.8373	< 0.001
Standard Deviation	0.8140	0.8384	< 0.001
Counts	0.8210	*0.8417*	–

6.4.2 Control and Basic Variables

Our first set of models seeks to test the difference between the basic models presented in [1] and the same models with clinical variables added. Our analysis yielded many candidate models. The worst of all the new models was that including the same variables as the original "both" model from Chapter 5 with the addition of the entropy of the medications received. Its average AUC was 0.8449 which still significantly outperforms (p-value < 0.001) even the best model from 5 and [1].

The results are in Tables 6.6, 6.7, and 6.8 and show that the lab results give the largest boost to model performance. The results are shown rounded to four decimal places in order to better distinguish between the models.

All of the models including laboratory test results were well above the previous best model. The worst performing model in this set is that which included the lab result indicator variable with an AUC of 0.8626 which is still just under 0.02 higher. The best clinical variable model in this set was that including the minimum lab result at 0.8677. The average AUC of the models is 0.8652 with a standard error of 0.0022.

The p-values from the aforementioned Wilcoxon signed rank tests in Table 6.6 are all relative to the model with the highest average AUC, Minimum Lab Result. The AUC for the minimum lab results was significantly greater than all the other models in this collection.

One notable result is the improvement of the models from the first run to the second run. The first run included all the variables and the second run reduced according to the procedure described in Section 8.3. The second run increased the AUC an average of 0.02218.

Table 6.6: Average AUCs for first and second runs (50 runs each) of each candidate lab result model. Each model included the Both dataset from [1] and the stated derived variables.

Model	Run 1	Run 2	p-value
Minimum Lab Result	0.8444	*0.8677*	–
Maximum Lab Result	0.8444	0.8670	< 0.001
Last Lab Result	0.8452	0.8666	< 0.001
Sum of Squared Lab Abnormality	0.8408	0.8645	< 0.001
Number of Labs	0.8394	0.8630	< 0.001
Lab Result Indicator	0.8441	0.8626	< 0.001

The receiver operating characteristic curves shown in Figure 6.1 and in all the figures in this section are representative of their respective fifty test set runs. The curves do

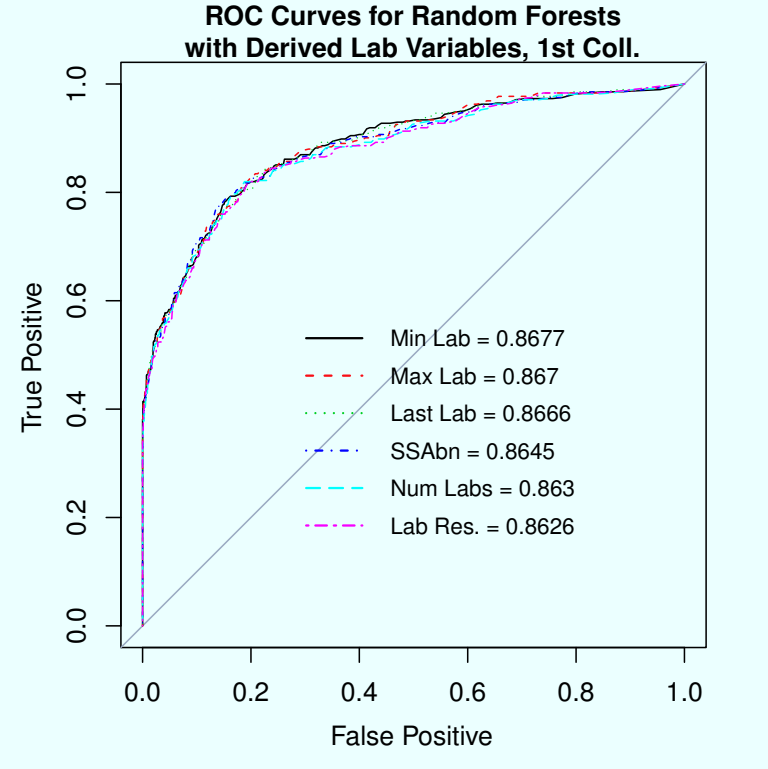

Figure 6.1: ROC curves for lab models in collection one.

not show any one model dominating the others. From 0.2 to 0.6 on the False Positive axis the lab result indicator seems to perform worse.

The models including metrics based on the medications administered during inpatient visits showed similar patterns. The best performing model among those including medication metrics was the gini metric of variation in the medication administration with an AUC of 0.8475. The worst model of the group was the entropy variation metric with an AUC of 0.8449. The average of all the medication model AUCs in collection 1 was 0.8462 with a standard error of 0.0009. The average increase between run 1 and run 2 was 0.0338. This was a larger increase than that seen for laboratory results.

The p-values in Table 6.7 all result from using the Wilcoxon signed rank test on the Gini index of medication counts against the remaining model AUCs. In each case

the difference was statistically significant with the lowest significance being against the maximum number of medications administered. The number of days a medication was given had a significant p-value of 0.001.

Table 6.7: Average AUCs for first and second runs (50 runs each) of each candidate medication administration model. Each model included the Both dataset from [1] and the stated derived variables.

Model	Run 1	Run 2	p-value
Maximum Number of Times Each Med Given	0.8133	0.8467	0.019
Total Visit Count of Each Med Given	0.8130	0.8454	< 0.001
Number of Days Each Med Given	0.8130	0.8462	0.001
SD of Daily Med Counts	0.8118	0.8463	< 0.001
Gini of Daily Med Counts	0.8117	*0.8475*	–
Entropy of Daily Med Counts	0.8113	0.8449	< 0.001

The corresponding ROC curves to the medication models in collection one shown in Figure 6.5 do not show any dominance. The gini model does tend to be higher from 0.2 to 0.4 on the false positive axis but no dominance is evident overall.

The best models of the lab results and medications administered were combined to create a new model with the hopes that combined they would produce a better model than each individually. As such we combined the minimum lab result and the gini coefficient of the medications to create the larger combined model. We see in Table 6.8 that the combined version of both models has a final AUC of 0.8649. As notable values of lab results can return low or high we also created a model with the minimum and maximum lab results. The final AUC of this model was 0.8674.

Table 6.8 shows a model which does not directly include derived lab variables. It is a model which included the base model with the control variables plus all of the aggregated variables described in Section 6.2.3 except the number of prior visits and thirty-day indicator. This aggregate model is much lower dimension than the other models and yet has a high AUC of 0.8587. In fact, it is higher than all of the medication

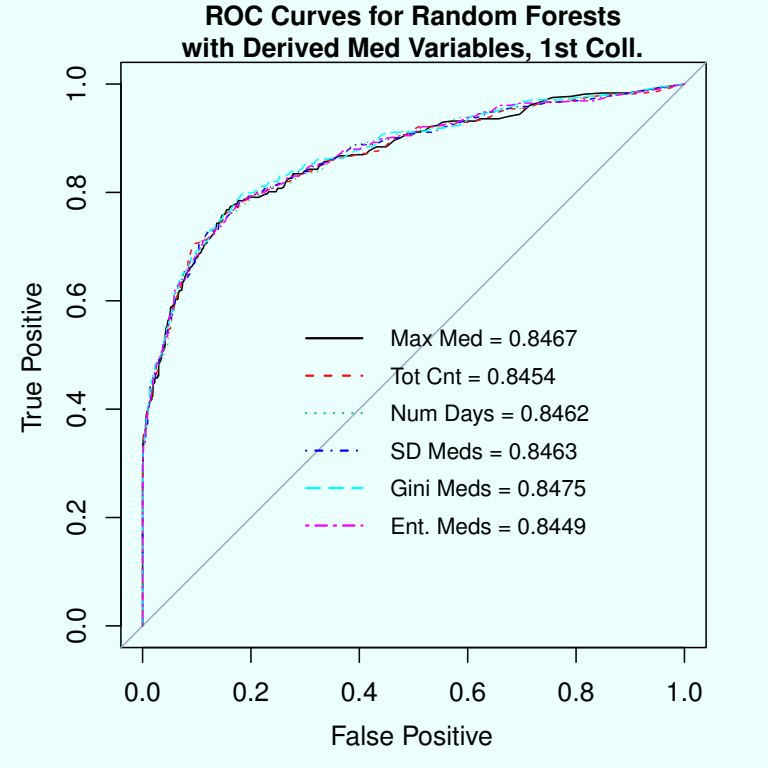

Figure 6.2: ROC curves for medication models in collection one.

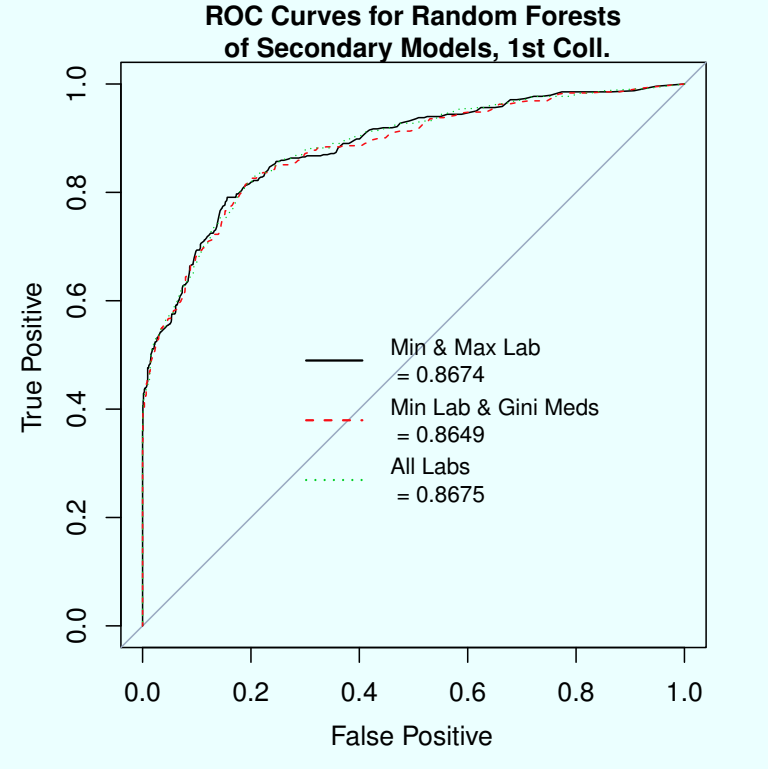

Figure 6.3: ROC curves for secondary models in collection one.

models.

Table 6.8: Average AUCs for first and second runs (50 runs each) of secondary models. Each model included the Both dataset from [1] and the stated derived variables.

Model	Run 1	Run 2	p-value
Base + Lab & Med Aggregates	0.8587	–	< 0.001
Minimum & Maximum Lab Result	0.8459	0.8674	0.462
Minimum Lab & Gini of Daily Med Count	0.8380	0.8649	< 0.001
All Derived Lab Variables	0.8447	0.8675	–

ROC curves for the secondary models with lab results in non-aggregated form in Figure 6.3 show that none of the models dominate.

However, when comparing the best lab model in collection one with the best medication model in collection one in Figure 6.4 the minimum lab model dominates the gini medication model for most of the curve with only a small overlap near 0.1 on the

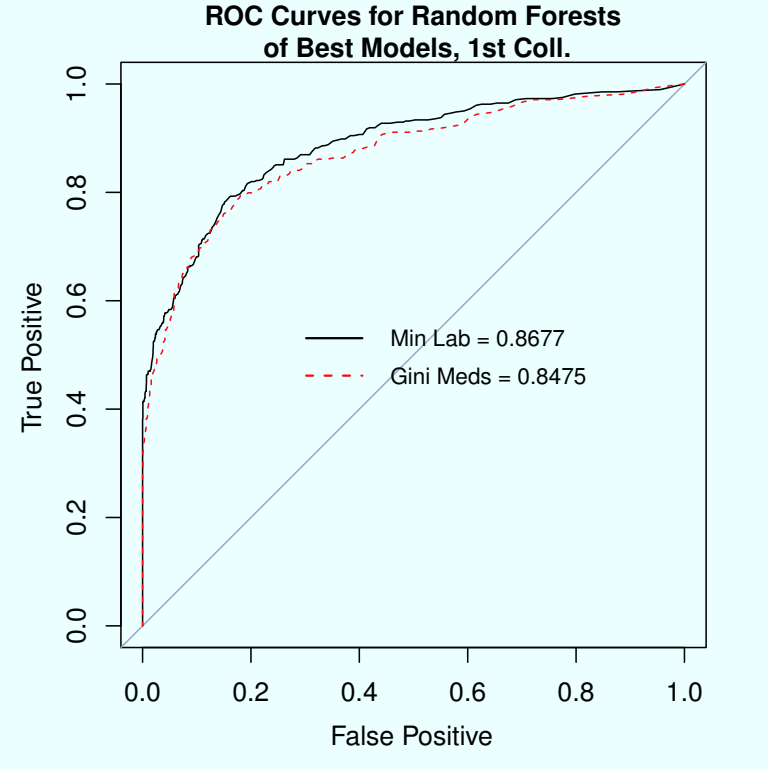

Figure 6.4: ROC curves comparing the best models in the lab and medication sub-collections in collection one.

false positive axis.

Variable Importances

The variable importances for the minimum lab results found in Table 6.9 show that the most important variable is the length of stay followed by 99232 which is a Level 2 Hospital Progress Note. The third most important variable is the payor which could be an indicator of socio-economic conditions. The first lab result is the ninth most important and is ALB which is albumin serum used to check nutritional status. The next lab result is the eleventh most important and is Na+ which is sodium.

Table 6.9: Table of top 25 most important variables for the best models.

Rank	Min Lab	Gini Med	Min & Max Lab	Min Lab & Gini Med
1	los	los	los	Payor
2	99232	99232	99232	Age
3	Payor	Payor	ALB.min	los
4	93010	93010	ALB.max	511.9
5	Age	Age	Payor	585.9
6	99233	99233	99233	401.9
7	ALB	514	93010	ALB.min
8	71010	71010	71010	99214
9	511.9	511.9	511.9	428.0
10	99214	428.0	Age	BNP.min
11	514	99214	99214	585.6
12	428.0	429.3	514	71010
13	585.9	71020	428.0	424.1
14	429.3	585.9	585.9	414.01
15	CREAT	99238	429.3	514
16	Na+	99223	99238	427.31
17	BUN	786.05	99223	Gender
18	71020	518.0	71020	250.00
19	CL	401.9	CREAT.max	429.3
20	CALCM	793.1	Na+.max	Race
21	99238	99255	Na+.min	BUN.min
22	GLUC	427.31	CREAT.min	99233
23	99223	585.6	BUN.min	414.00
24	K	99291	CL.max	599.0
25	CO2	99254	BUN.max	425.4

6.4.3 Control, Basic, Derived, and Clinical Variables

The results in this section reflect the addition of two variables not reported in [1]: number of prior inpatient admissions and an indicator of whether the current visit is itself an unplanned thirty-day readmission. These aggregated metric variables are discussed in Section 6.2.3. The best laboratory result based model in Table 6.10 in collection two is that based on the maximum lab results with an AUC of 0.8719.

The p-values in Table 6.10 all result from using the Wilcoxon signed rank test on the

maximum lab result model against the remaining model AUCs. In all but one case, the difference was statistically significant with an $\alpha = 0.001$. The model had significantly higher AUCs with an $\alpha = 0.1$ when compared to the minimum lab result set.

The worst model of this particular collection is the sum of squared abnormality with an AUC of 0.8671. The average AUC of all the AUCs in this collection is 0.8699, just under 0.87, with a standard error of 0.0019. The average increase from run 1 to run 2 was 0.0181.

Table 6.10: Average AUCs for first and second runs (50 runs each) of each candidate lab result model with additional summary variables. Each model included the Both dataset from [1] and the lab result derived variables.

Model	Run 1	Run 2	p-value
Minimum Lab Result	0.8533	0.8716	0.087
Maximum Lab Result	0.8536	*0.8719*	–
Last Lab Result	0.8534	0.8701	< 0.001
Sum of Squared Lab Abnormality	0.8494	0.8671	< 0.001
Number of Labs	0.8482	0.8680	< 0.001
Lab Result Indicator	0.8526	0.8704	< 0.001

Like the models from the first collection, Figure 6.1 shows that no model dominated any other for the second collection lab models.

Table 6.11 shows that the best model of the medication models in this collection is the total visit count of each medication given with an AUC of 0.8540. The worst performing model was the standard deviation of the daily medication counts with an AUC of 0.8490. The average AUC for this model collection was 0.8508 with a standard error of 0.0019. The average increase in AUC from the first to the second run was 0.0300.

The p-values in Table 6.11 all result from using the Wilcoxon signed rank test on the total visit count of each medication given model against the remaining models' AUCs. In each case the difference was statistically significant with an $\alpha = 0.001$.

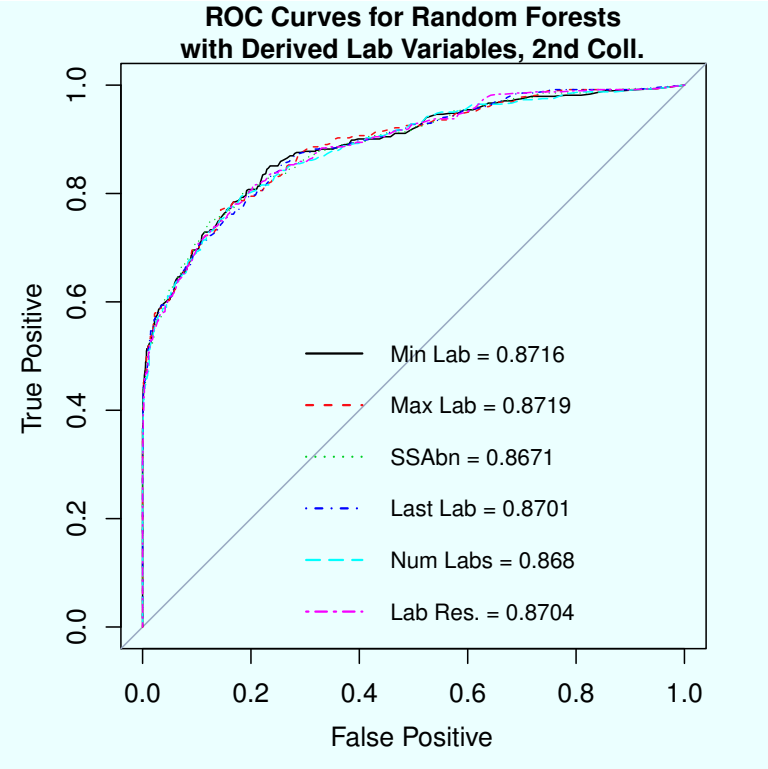

Figure 6.5: ROC curves for the lab models in collection two.

Table 6.11: Average AUCs for first and second runs (50 runs each) of each model with derived medication variables and clinical variables. Each model included the Both dataset from [1] as well as novel derived variables, and summary variables.

Model	Run 1	Run 2	p-value
Maximum Number of Times Each Med Given	0.8217	0.8506	< 0.001
Total Visit Count of Each Med Given	0.8215	*0.8540*	–
Number of Days Each Med Given	0.8215	0.8493	< 0.001
SD of Daily Med Counts	0.8200	0.8490	< 0.001
Gini of Daily Med Counts	0.8204	0.8518	< 0.001
Entropy of Daily Med Counts	0.8201	0.8503	< 0.001

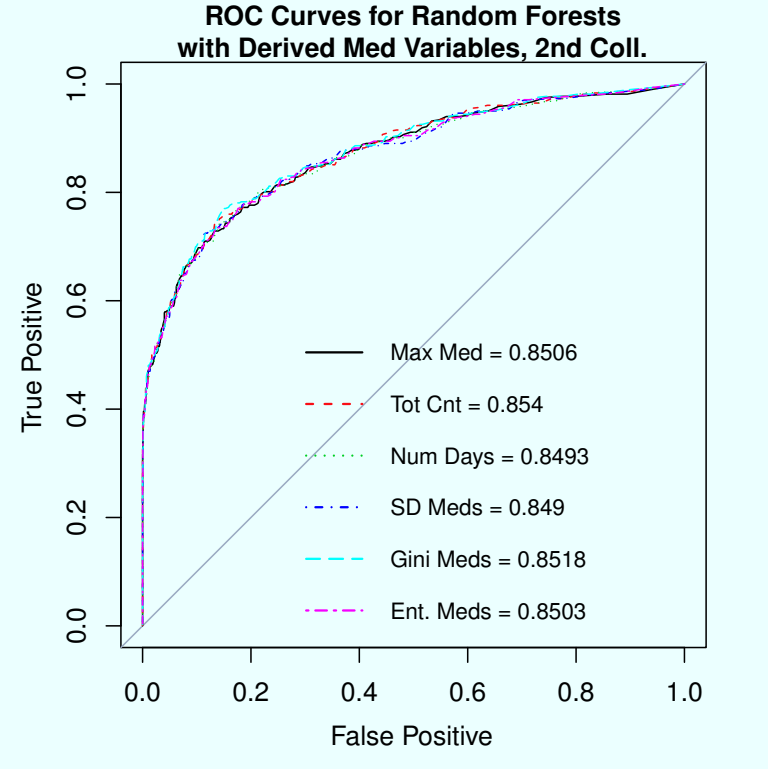

Figure 6.6: ROC curves for the medication models in collection two.

We see in Figure 6.6 that the pattern of non-dominance within a particular type of model (E.g., within lab based models, or within med based models) continues.

The secondary models in Table 6.12 include a model where the best models from the lab results type and the medication type were combined to make a model with AUC 0.8722. Because the medication models did not seem to add as much information to the models the total visit counts were aggregated across all visits and appended as a single variable to the maximum lab results variables to form another model. The AUC for the simplified combined model was 0.8711. The base model with the aggregates also represented. The only difference between this base and aggregates model and the one presented in Table 6.8 is the addition of the two new aggregated metrics, number of prior visits and the thirty-day indicator variable.

Table 6.12: Average AUCs for first and second runs (50 runs each) of secondary models that include the new derived variables. Each model included the Both dataset from [1] as well as the number of prior inpatient visits and variable indicating whether the current visit is itself a 30-day readmit as well as the stated variables.

Model	Run 1	Run 2	p-value
Base + Lab & Med Aggregates	0.8651	–	< 0.001
Max Lab Result & Tot. Vis. Cnt. of Each Med Given	0.8477	*0.8722*	–
Max Lab & Sum of Tot. Vis. Cnts. for All Meds Given	0.8534	0.8711	< 0.001

The AUCs of the combination model of max lab and total visit count of each medication results were significantly greater than other two models in the table with $\alpha = 0.001$. The secondary model ROCs for the second collection as seen in Figure 6.7 also showed no dominance between the complex models (i.e., the models that weren't only aggregate variables).

The medication total count model in Figure 7.4 is dominated by the lab-based models. There does not appear to be any dominance among the lab-based models in collection two.

Variable Importances for Full Models

Table 6.13 shows the variable importance rankings for the models including the two additional aggregated variables. The top four variables across each model are the number of prior visits, the payor, the age and the length of stay in that order. The fifth and sixth ranks have 401.9, and 511.9 as well as 414.01 which are hypertension, unspecified pleural effusion, and coronary atherosclerosis of a native coronary artery respectively.

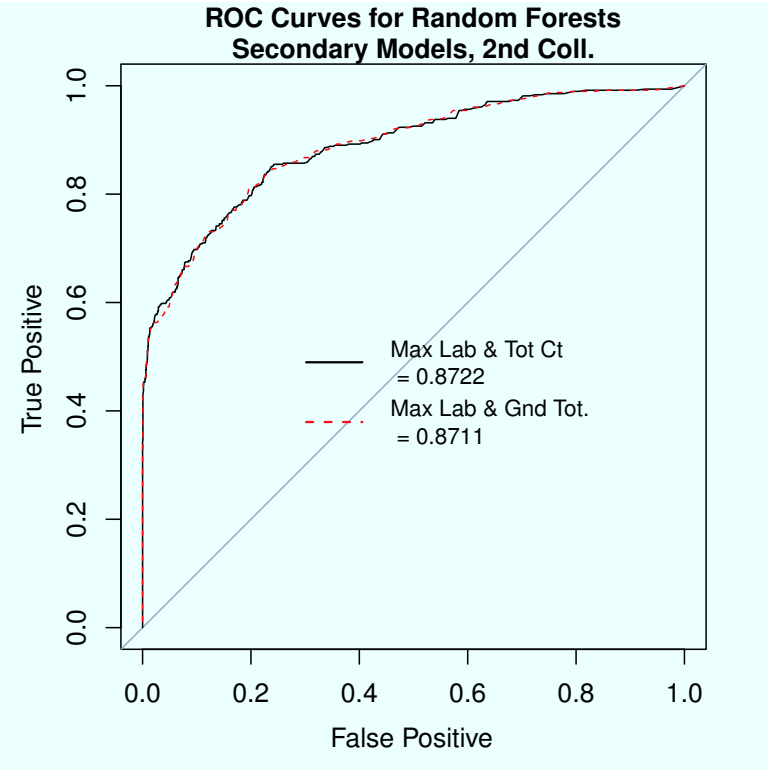

Figure 6.7: ROC curves comparing the best overall model in collection two to the best lab-based model in collection two.

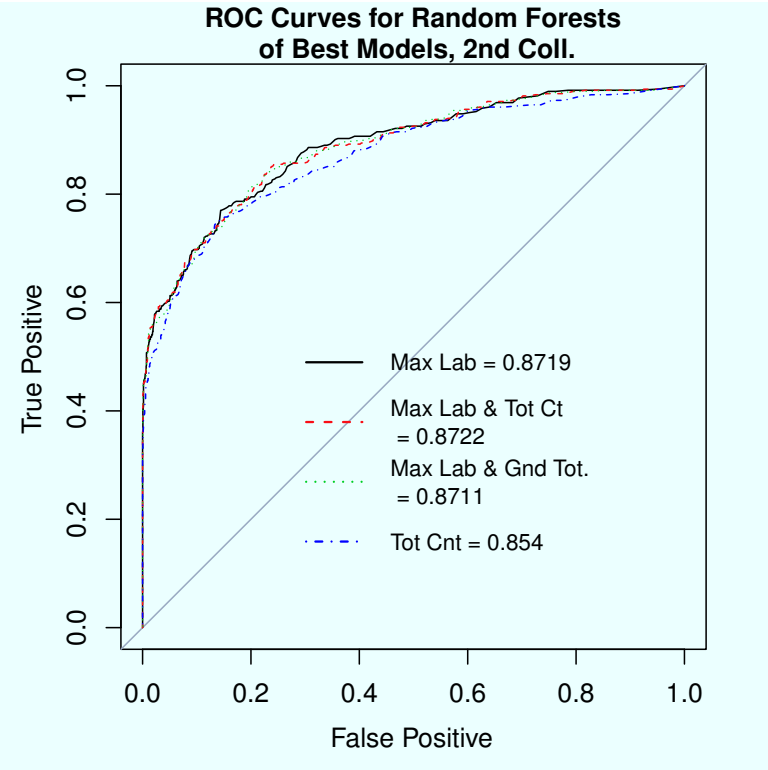

Figure 6.8: ROC curves comparing the best lab-based models and the best medication model in collection two.

Table 6.13: Variable importances for second collection of models. The top 25 for each model is shown. This is the ranking averaged over the 50 runs for each model.

Rank	Max Lab	Total Meds	Max Lab & Tot. Meds	Max Lab & Grand Tot. Meds
1	numPriorVisits	numPriorVisits	numPriorVisits	numPriorVisits
2	Payor	Payor	Payor	Payor
3	Age	Age	Age	Age
4	los	los	los	los
5	401.9	401.9	511.9	401.9
6	414.01	511.9	401.9	511.9
7	511.9	414.01	585.9	414.01
8	585.9	250.00	ALB.max	250.00
9	427.31	428.0	99214	428.0
10	BNP.max	427.31	428.0	427.31
11	Race	Race	BNP.max	Race
12	250.00	Gender	414.01	Gender
13	BUN.max	585.9	585.6	585.9
14	Gender	99214	ThirtyDayCase	99214
15	414.00	93010	71010	93010
16	ALB.max	425.4	424.1	425.4
17	424.1	414.00	427.31	414.00
18	425.4	786.05	514	786.05
19	428.0	429.3	Gender	429.3
20	496	585.6	99232	585.6
21	99214	427.32	BUN.max	427.32
22	585.6	496	250.00	496
23	CREAT.max	514	Race	514
24	ThirtyDayCase	599.0	414.00	599.0
25	93010	424.1	429.3	424.1

Comparisons Across Both Collections

It is useful to compare the models that have different meaning or variable bases from both collections one and two. In Table 6.14 we have representatives from both collections. The table shows that the overall most predictive model as measured by AUC is from the second collection which included the max labs and the sum of total visit counts for each medication with AUC 0.8722. It is significantly greater than the simpler

max lab (of collection two) with a p-value of 0.059.

Table 6.14: Comparison of final selection of models.

Model 1	AUC	Model 2	AUC	p-value
Base, Lab & Med Agg. 2	0.8651	Base, Lab & Med Agg. 1	0.8587	< 0.001
Min Lab Coll. 1	0.8677	Base, Lab & Med Agg. 1	0.8587	< 0.001
Min Lab Coll. 1	0.8677	Base, Lab & Med Agg. 2	0.8651	0.067
Min Lab Coll. 1	0.8677	All Labs Coll. 1	0.8675	0.343
Max Lab 2	0.8719	All Labs Coll. 1	0.8675	< 0.001
Max Lab 2	0.8719	Min Lab Coll. 1	0.8677	< 0.001
Max Lab 2	0.8719	Max Lab 2 & Sum of Tot. Vis. Cnts. for *All* Meds	0.8711	0.002
Max Lab 2 & Sum of Tot. Vis. Cnts. for *Each* Med	**0.8722**	Max Lab 2	0.8719	0.059

The base and aggregate variables model from collection two outperforms collection one's (p-value < 0.001) but is in turn outperformed by the minimum lab model from collection one (p-value < 0.001). The minimum lab model outperforms the base model from the second collection (p-value$= 0.067$) but cannot significantly outperform the model from collection one which contains all of the lab variable types (p-value $= 0.343$). However, the max lab from collection two is able to outperform both the minimum lab from collection one (p-value < 0.001) and the model with all lab variables in collection one (p-value < 0.001). The max lab model also outperforms the combination model of itself with the variable that sums the visit counts for all medications (p-value $= 0.002$). In the end, it is outperformed by itself combined with the sum of total visit counts of each medication (p-value $= 0.059$).

6.5 Discussion

6.5.1 Indicator, Variation, and Counts Comparison

Each of the initial models based on the four sets of variables from Section 6.4.1 indicates something different. The indicator variables on procedures are a measure of risk associated with that procedure. The variation metrics each show that variation in procedure application is predictive of unplanned thirty-day readmissions. In fact, Table 6.5 shows that of the four variable types tested, the standard deviation variation metric has the best performance of the variation metrics and outperforms the indicator variable. We confirmed this using the usual test and correction which yielded a p-value of 0.059 which is significant when $\alpha = 0.1$. This provides stronger evidence than that found in Chapter 4.

While we found that counts of each procedure and diagnosis were more predictive in the end, we were able to show that within variation metrics were predictive of unplanned thirty-day readmissions. This implies that when we are interested in using them for predicting outcomes that reflect quality of care, they may be useful variables.

6.5.2 First Model Collection

It is clear from the tables and from Figure 6.4 that all derived lab result variables are more predictive than the derived medication variables. Though the minimum lab result significantly outperforms the other lab models it isn't an enormous performance jump at just 0.0007 greater on average. However, over enough patients such a difference becomes meaningful. Interestingly the sum of squared abnormality does not perform

best, even though it is designed to directly capture how much each patient is out of the normal range across a visit. The two top metrics are the extrema, which reflect the worst conditions of a patient, conditional on the lab test. Whether the minimum or the maximum performs better could have to do with which tests reflect worse conditions in each direction. The worst of the group is the generalized lab result indicator. Perhaps it aggregates the information too much whereas the other variables still offer some indication of degree of severity or extremity.

It isn't a surprise that the lab variables do better than the med variables as they directly measure the physiological state of the patients, whereas medications administered reflect physicians' attempts to address revealed symptoms. The medications are separated from the actual condition by many steps of information conveyance. Despite their lower performance, even the entropy metrics, the worst performing medication variable set, outperform the base model from Chapter 5 (p-value < 0.001). This provides evidence for the value of variation metrics as a source of information regarding patient care, and it also provides evidence that medication variables are useful in improving readmissions prediction.

Of the variation metrics, the Gini index of the daily medications administered performed the best, in fact, it outperformed all of the other metrics in this medication set. This provides stronger evidence of the merits of measuring variation of care, in this case by regularity or consistency of medications administered.

The ROC curves demonstrate the lack of dominance in both the lab and med groups, and so the AUC comparisons are best for measuring overall performance. The ROC curves are helpful in identifying whether any particular models are better classifiers of one or the other class. It appears all seem to perform about the same in classifying both classes. All of the models have a bit more difficulty classifying the readmits which

is reasonable as they are the smaller class, though they still do well.

Somewhat surprisingly, none of the secondary models for the first collection outperformed the minimum lab model. This could be due to the inability of the random forests to sort through the number of noisy variables present in the much larger variable sets. The model that did very well considering the variables it was built on was the base model with aggregated lab and med variables. With an AUC of 0.8587 this model is much better than the standard both model (p-value< 0.001). It does worse than the model containing all of the lab variables together, but considering its parsimony and arguably better interpretability, it may be a good choice for implementation in some circumstances.

The variable importances for the best models show remarkable similarity between the min lab and gini med models. The top six most important variables are the same for each. Notably none of the top twenty-five important variables are medications administered. This could explain why medication based models did much worse than the lab models. It could also explain why the combined lab and med model did worse than the lab alone. While they add information to the original both model, the med variables appear to be much noisier than the lab variables.

The albumin serum appears to be the most important of the lab variables in all of the models with lab results in them with creatinine appearing in two of them. The payor variable is in the top three in three of the models and is in the top five of all of them. This reflects a similar finding to Amarasingham et al. [127] that the mode of payment was important. They suggested it may reflect socioeconomic status which could be a proxy for many things. The 99232 procedure code indicates subsequent hospital care and could be another way of measuring how long their stay was. Of course, the length of stay (los) is the most direct measure of that. The diagnosis

code 514 (pulmonary congestion and hypostasis) appears in the top fifteen of all four. These variable importance rankings for the various models all provide a springboard of inquiry into what may be related to unplanned readmissions.

6.5.3 Second Model Collection

The second set of models, which include two additional non-medical variables over the first collection, follows a similar pattern to the first collection. Again, the lab result based models all outperform the medication based models by a fair margin. The top two lab result models were again the min and max lab results, however with the additional two variables, the maximum lab result outperformed the minimum lab result though significant only to $\alpha = 0.1$. The max lab model was significantly better than all of the rest with an $\alpha = 0.001$ This round the lab result indicator faired much better in third place with the last lab in fourth. The sum of squared abnormality still performed the worst, a bit behind the number of labs metric. In all cases, they outperform the both model. The ROC curves in Figure 6.5 show overlap among all of the lab models. Again using tests on the AUCs seems appropriate for ranking the models' performance.

The medication based models in the second collection had an interesting change in rankings. The total visit count of each medication given went from fifth of six to the top performing model after the addition of the two new variables. It could be there is some sort of interaction that would produce such a drastic change in rank. Though it did displace the Gini index as the best model, the Gini model was still the second best. The third best was the maximum number of times a medication was given with the entropy close behind. Again the variation metrics on medication administration perform well when used as variables to predict thirty day readmissions.

Only one secondary model outperforms the best from the lab and med models in collection two. When the best lab and med models are combined they form a model that outperforms all of the others so far. It seems the addition of the two new derived variables provide enough information to overcome the noise in the larger datasets. The base model performs very well with the additional two variables and outperforms all of the medication based models. This is an impressive result as the variables are much more aggregated and separated in terms of the information they convey about the state of the patients. The aggregate model is essentially a collection of counts of various models which could be seen as indicators of illness severity, but the exact condition or reason is unknown. Because collection one showed that the med results could sometimes reduce model performance when combined, we built a combined model with the total medication counts aggregated across the entire visit. This aggregation still did not help as it brought the max lab model down from 0.8719 to 0.8711. Interestingly, with the presence of the two additional variables, having the count variables "expanded" improves the performance unlike in collection one. Figure 7.4 shows that the best models with lab results do not dominate one another but all three appear to dominate the best medication model.

The variable importances of these models are very similar. The two four variables are all identical across the best models with the fifth and sixth variables having switched in a couple models. The most important variable across the board is how many prior inpatient visits a patient had. This could indicate how chronically ill a patient is which would intuitively make sense to predict whether a patient would come back, and soon. The second most important variable was the payor, which as stated before, could indicate socio-economic status. This could proxy for many things such as access to quality care (outside of the hosptial) as well as education, health education and others. This is clearly conjecture but would make for an interesting study. Age is the

third most important and could indicate frailty and robustness. The length of stay is the fourth and may indicate how ill a patient is on this particular visit. The fifth has 401.9 which is hypertension. Obviously hypertension is on clinician's radar already but the striking agreement between the models is impressive. Diagnosis 511.9 is unspecified pleural effusion, or fluid in the chest.

Type II Diabetes (250.00) shows up in the eighth slot for both the med model and the max lab with the aggregated med model. As for important lab tests albumin shows up as well as B-type Natriuretic Peptide (BNP, indicates when heart failure worsens). Gender also is in the top 25 of all four models with the highest rank at twelfth.

In the model with the grand total of medications administered per visit there are no lab results in the top twenty-five variables. This may suggest that rather than the medication variable interfering with the lab results, the lab results may be interfering with the medication variable. However, there are also no medications in the top twenty five variables. The consistency in ranking among the top models in collection two could provide guidance for investigators looking into what may be causing or strongly related to unplanned thirty-day readmissions.

6.5.4 Overall Comments

Table 6.14 shows a progression of performance increases to the best overall model. The best overall model is from collection two (meaning it had the two additional variables) with both the max lab variables and the sum of of total visit counts for each medicine. This model significantly ($\alpha = 0.1$) outperformed the next best (max lab) model. Model selection based on pure performance would lead to this model. However for parsimony the aggregated model from collection two does very well. For the middle ground, the

max lab model would be a strong candidate.

6.6 Conclusions

The initial set of models showed that including both explicit clinical variables (laboratory results) and implicit clinical variables (medications administered) significantly improve predictive performance of unplanned thirty-day readmissions. We also found that adding two non-medical derived variables improved the models significantly. Variable importance rankings by select models highlighted some consistent patterns which may be useful for those with a medical background to further investigate. While these models are not causal, they do provide a ranking mechanism with some interpretability. Clinical variables both explicit and implicit when represented by appropriately chosen metrics along with well chosen non-medical variables form highly predictive Random Forests models.

Chapter 7

Thirty Day Readmissions Prediction Including Clinical Data using SVMs

Support vector machines have been shown to be powerful classifiers. This chapter compares the performance of support vector machines to the performance of the Random Forests models in Section 6.4.3. This chapter builds on the work from collection two in Chapter 6 by creating support vector machines for comparison with collection two's results. Additional select models are created and compared. The variables in this chapter are the same as Chapter 6. Performance results in the models are presented, compared, and discussed.

7.1 Background

Some have found that support vector machines (SVMs) do not perform well on high dimensional data [181,197]. The authors refer to data from the "omics" which are very high dimensional with tens of thousands of variables but perform their simulation on a dataset with 1,000 variables. The data for this research include thousands of variables with some models containing about 20,000 variables (all lab results model). Hastie et al. [150] suggest that SVMs have a number of limitations that may prevent them from doing well on sparse high dimensional data (see chart on page 351 of their book). Included in these suggestions is the idea that SVMs cannot handle irrelevant inputs very well. The data for this dissertation includes many sparse variables whose relevance is difficult to determine. Based on results from the prior chapter, the medical data is less relevant than the lab data, though not completely irrelevant as a whole. Other criticisms include their inability to handle categorical variables in a "natural" way, as well as missing values, both of which are present in this data. Another weakness of SVMs is their inherent lack of interpretability. The studies currently available using SVMs have not found marked improvement in predicting unplanned thirty-day readmissions over logistic regression. As we found excellent results for Random Forests where none was found in prior studies, we also explore the suitability of SVMs using our new variable creation and model selection paradigm.

7.2 Methodology

Support vector machines were run on the same variable sets as in Section 6.4.3. We optimized the parameters using a grid search with 10-fold cross-validation over two kernels: a linear kernel, and a radial basis kernel using the `tune.svm()` function [198]

in R. The optimal setting for the linear kernel was $cost = 2$. The optimal radial basis function settings were $\gamma = 0.0625$ and $cost = 1$. Of the two, the best performing kernel was the radial basis function (RBF). We compared performance for final parameter selection using AUC on the lab results variables. The linear kernel's AUC was 0.7432 as opposed to the RBF's 0.882. Because of its poor performance, the linear kernel was no longer used and the rest of the analysis is performed using the RBF kernel with stated parameter values. As already established for other algorithms, the metric used for model selection is area under the curve or AUC.

7.3 Results

The support vector machines predicting using lab result metrics in Table 7.1 all outperformed their Random Forests counterparts. The p-values in Table 7.1 all result from using the Wilcoxon signed rank test comparing the SVM model with the Random Forest model. The test in each case was that the location shift of the SVM AUCs is not greater than zero when compared with the corresponding RF model. In each case the difference was statistically significant with an $\alpha = 0.1$. The p-values for the sum of squared abnormality, number of labs performed, and the lab result indicator were all < 0.001. The minimum lab result was significant at a 95% confidence level. The best performing model overall is the lab result indicator using SVMs with an AUC of 0.882. This is just over 0.01 higher than the best model in the RF group.

In Figure 7.1 we can see that for much of the false positive range the two models perform similarly. The range where the SVM model seems to dominate is from 0.4 to 0.6, which can be a tricky area and also where the most natural decision boundary, 0.5, is.

Table 7.1: Average AUCs over 50 runs for the Support Vector Machine and Random Forests lab result models. The variables are the same as the lab result models in Section 6.4.3.

Variables	SVM	RF	p-value
Minimum Lab Result	0.8727	0.8716	0.035
Maximum Lab Result	0.8721	0.8719	0.093
Last Lab Result	0.8723	0.8701	0.005
Sum of Squared Lab Abnormality	0.8725	0.8671	< 0.001
Number of Labs	0.8815	0.8680	< 0.001
Lab Result Indicator	0.882	0.8704	< 0.001

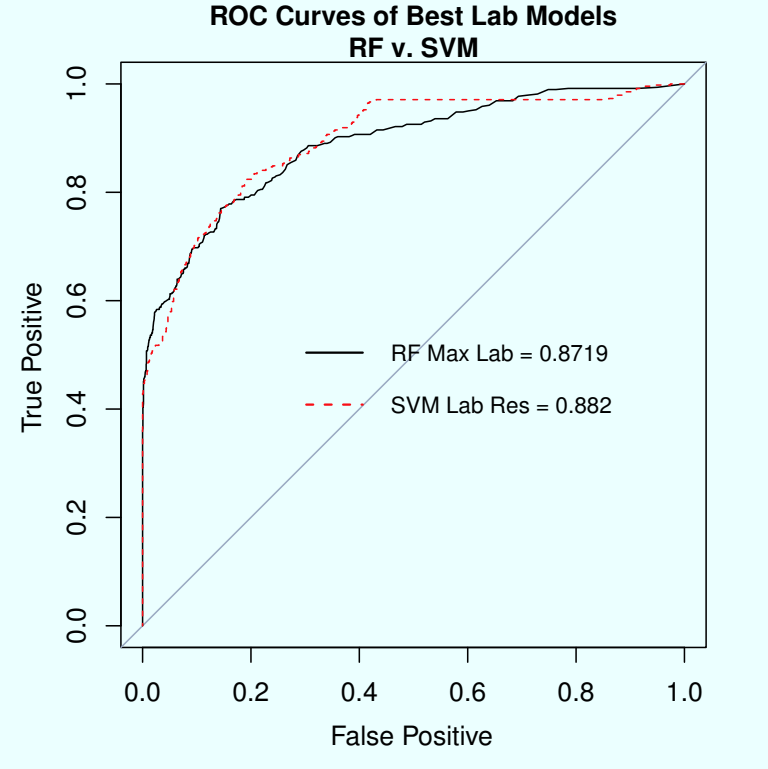

Figure 7.1: ROC curves for the best lab models between Random Forests and SVMs.

For the medication based models, we see in Table 7.2 that the SVM models handily outperformed their random forests counterparts. The best performing model overall was the SVM model utilizing the maximum number of times a medication was used metric, with an AUC of 0.8707. The worst performing model of the SVM group was a three way tie amongst the variation metrics with AUCs of 0.8698. Even the worst model outperformed the best of the Random Forests group. The p-values in Table 7.2 all result from using the Wilcoxon signed rank test comparing the AUCs of the SVM to its corresponding RF model. In each case the difference was statistically significant with an $\alpha = 0.001$.

Table 7.2: Average AUCs for the support vector machine and Random Forests (50 runs each) of each candidate medication model with additional summary variables. Each model included the Both dataset from [1] and the lab result derived variables.

Variables	SVM	RF	p-value
Maximum Number of Times Each Med Given	0.8707	0.8506	< 0.001
Total Visit Count of Each Med Given	0.8705	0.8540	< 0.001
Number of Days Each Med Given	0.8706	0.8493	< 0.001
SD of Daily Med Counts	0.8698	0.8490	< 0.001
Gini of Daily Med Counts	0.8698	0.8518	< 0.001
Entropy of Daily Med Counts	0.8698	0.8503	< 0.001

In Figure 7.2 we compare the best SVM model to the best RF model for the medication group. The SVM model dominates the other from 0.1 to 0.6 on the false positive range and is about the same over the rest of the range. The SVM model does seem to dip below the RF model on both this and Figure 7.1 for a small range around 0.75. This dip is much smaller than SVM's dominating range for both variable groups.

We also compare select secondary models. We see in Table 7.3 that the overall best model is in the SVM group with an AUC of 0.8827. This particular model was the lab result combined with the maximum medication models. The best model for the RF group was its best which was the maximum lab with the total visit counts for all meds

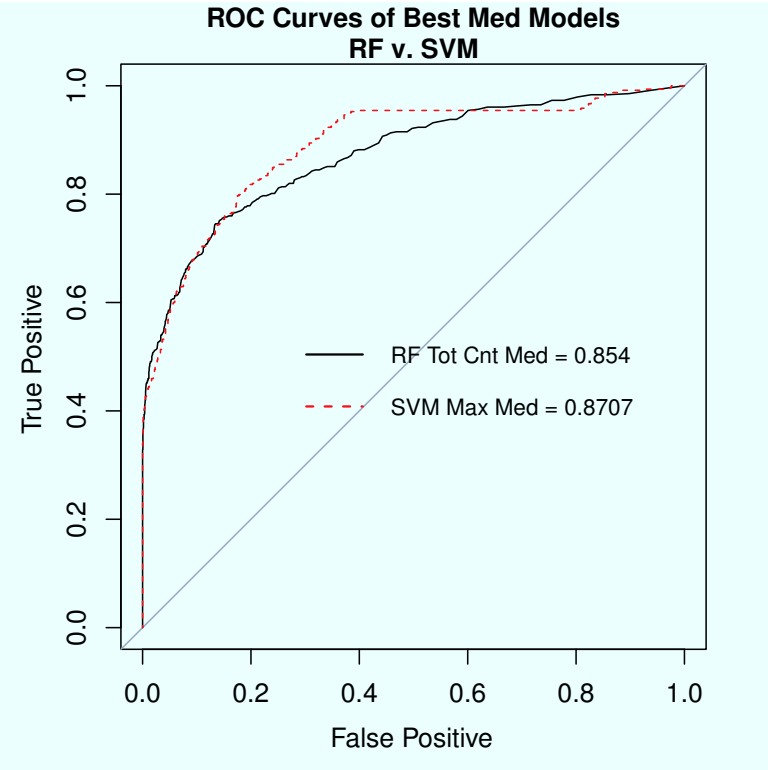

Figure 7.2: ROC curves for the two best medication models for each of the SVM and RF groups.

given with an AUC of 0.8722, as already reported. We also compared the performance of the two algorithms on the aggregated variables. This was the only model out of the entire series where the SVM underperformed the Random Forests. Not only did it underperform but drastically so with an AUC of 0.6975, while the RF had an AUC of 0.8651. The SVM version of the best overall model from the random forest collection had a higher average AUC but was only significantly better with an $\alpha = 0.15$. The two algorithms performed similarly when all the lab metrics were used with SVMs besting RFs by a statistically significant p-value= 0.005 but not materially significant 0.002 (0.8695 vs. 0.8675). When combining the best lab metric from each algorithm with the total visit counts for all meds given (aggregated total visit counts) the SVM version handily bested its counterpart with an AUC of 0.8824. This was only slightly lower than SVM's best overall model.

Table 7.3: Average AUCs for the Support Vector Machine and Random Forests (50 runs each) of secondary models.

Variables	SVM	RF	p-value
Base + Lab & Med Aggregates	0.6975	0.8651	1
Max Lab Result & Tot. Vis. Cnt. of Each Med Given	0.8724	0.8722	0.123
Best Lab & Sum of Tot. Vis. Cnts. for All Meds Given	0.8824	0.8711	< 0.001
All Labs	0.8695	0.8675	0.005
Best Model	0.8827	0.8722	< 0.001

The ROC curves of the models with the highest average AUC from Table 7.3 are shown in Figure 7.3. The pattern displayed in the earlier ROC figures continues in this group as well. The SVM outperforms in the middle of the false positive range then dips below around 0.8.

Figure 7.4 shows the incredible underperformance of the SVM algorithm on the aggregated variables. The RF version clearly dominates throughout the entire false positive range.

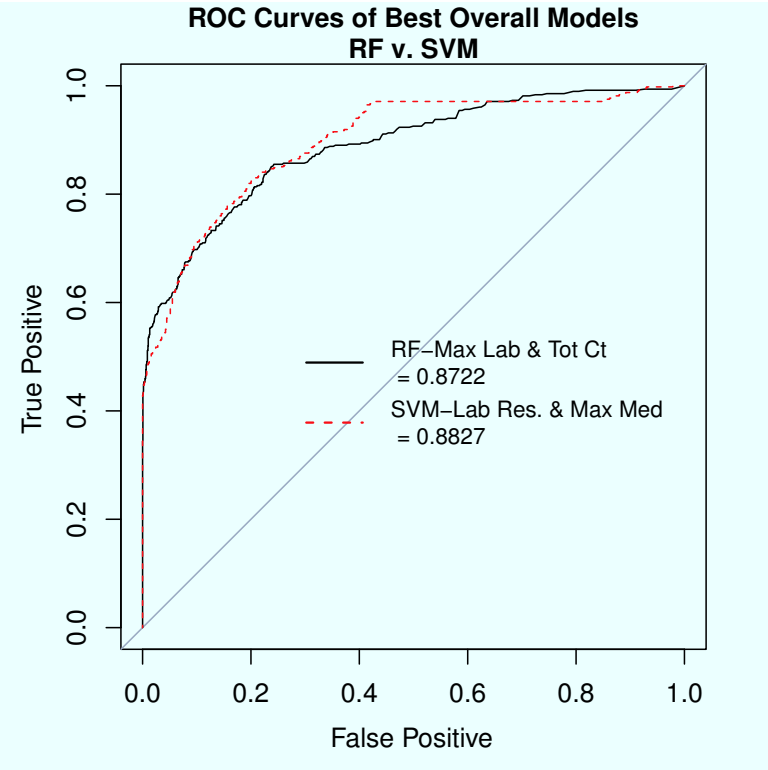

Figure 7.3: ROC curves comparing the best overall model in collection two to the best lab-based model in collection two.

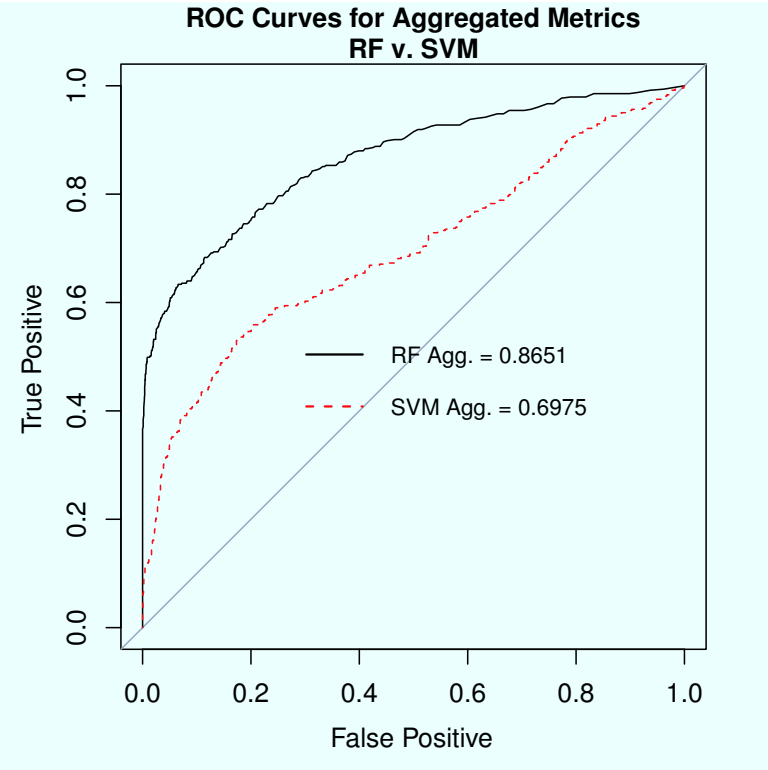

Figure 7.4: ROC curves comparing the best lab-based models and the best medication model in collection two.

Overall the SVM models outperformed the Random Forests models with the exception of the aggregated metrics model.

7.4 Discussion

The support vector machine results across both variable types are quite remarkable in light of prior statements on their suitability for such data. The SVM models absolutely dominate the Random Forests models on high-dimensional data in terms of AUC. What's more is which models perform well. The worst performing variable in the Random Forests models, sum of squared abnormality, became the third highest performing model when SVMs were used. The best model for the lab results models was the lab result indicator variable, which also did not perform as well in the Random Forests models. Interestingly, the best performing Random Forests (RFs) variable set, max lab, was the worst performing SVM model, although it still nominally did better than its Random Forests counterpart. It's almost as if the SVM results are a quasi-reverse ordering of the Random Forests models in terms of rank, but all outperforming the RFs.

If the lab results weren't astounding then the medical results could be as all of them performed comparably to the laboratory based RF models. In the SVM models, the variation metrics all had very similar performance, when rounding to the 4th digit, were the same. They also happened to not do quite as well as the count based variables, though nominally the difference is not large, though statistically significant. This result provides much food for thought. Are medication variables irrelevant or did the RFs just handle them poorly? Select models from the SVMs and RFs for each type of variable show a pattern in prediction for the models. Figure 7.1 and 7.2 show that the

SVMs dominate for much of the range then slip slightly under the RF curves in the higher false positive range.

Secondary models also showed that SVMs improved performance when best models from the lab and med groups were combined. The best lab for the SVMs was the lab result indicator and it was combined with the total visit count of the medications variable. When compared with the RF's best lab with the same med variable group, SVMs were nominally better though not significantly ($\alpha = 0.1$). Figure 7.4 shows that the simple aggregate variables that had performed so well with RFs did very poorly under SVMs. This may raise questions as to SVMs suitability for very-low dimensional datasets as well.

One of the most drastic differences of performance was in the medication group of models where the worst SVM model still performed as well as many of the RF lab models. If we use our medication results from Random Forests as evidence for the med variables' relative "irrelevance" then our SVM results may well provide evidence contrary to some of Hastie et al.'s suggestions. Support vector machines also did very poorly, relative to Random Forests, on the smaller models using the aggregated variables. This would also seem to contradict the authors' suggestions.

Perhaps part of the issue is a lack of clarity about what "high dimensional" data means, and what "irrelevant variables" actually are. Seeing as Hastie et al.'s suggestions are merely guidelines and not hard and fast rules, and that the data used here is distinct from that of [181, 197] we may be shedding a bit more light on those definitions. If SVMs truly don't do well on high-dimensional data, then perhaps our billing and clinical data, for the sake of clarity on this point, is *nearly*-high-dimensional. Despite their strong performance on so many of the models, SVMs did not perform well when all labs were combined. Perhaps this may help indicate the line between

nearly-high and high-dimensional. It could also be due to many correlated variables being present and so be a reflection of many irrelevant variables. We have been able to show, that for this data at least, SVMs have not performed well on information-dense, low-dimensional data. So then a tentative conclusion could be that SVMs perform well on nearly-high-dimensional data with sparse variables and limited relevance. Or, if we consider our data truly high-dimensional, the conclusion could be that SVMs *do* perform well on high-dimensional data.

7.5 Conclusions

Support vector machines significantly and meaningfully outperform Random Forests on all variable types except the simple aggregates. We have shown that SVMs either contradict earlier findings of poor performance on high-dimensional data or provide evidence for the line between nearly-high-dimensional data and actual high-dimensional data. We have also shown that SVMs perform very well on sparse, arguably irrelevant, data when predicting unplanned thirty-day readmissions. SVMs handily outperform nearly all models from prior chapters. The best overall model was the support vector machine using a radial basis kernel function on lab result indicator variables combined with the total count of each medication given for a visit.

Chapter 8

Assessing Sampling Schemes Impact on Thirty Day Readmissions Prediction

The limited previous studies available have been inconclusive about whether sampling improves thirty-day readmission predictive performance. This chapter builds on the work from Chapters 6 and 7 by comparing select models' performance when various sampling schemes are used. The sampling schemes assessed are undersampling, oversampling, and SMOTE. Performance results in the models are presented, compared, and discussed.

8.1 Background

For some supervised learning algorithms and some datasets, class imbalance may reduce the predictive performance of models built on them. The dataset utilized for

readmissions prediction may fall prey to these issues as it is imbalanced. More specifically, 20.4% of visits result in readmissions while the other four-fifths do not result in early unplanned readmissions. In [1] we used cost penalties in the Random Forests models to account for class imbalance and found that they did not improve overall performance. Upon presenting these results in our conference presentation for the paper, it was suggested that we compare our results to those when sampling schemes are used. To address this suggestion for handling imbalance we applied simple undersampling, simple oversampling, and SMOTE to the datasets used for our analyses thus far.

8.2 Data

The data used in this chapter is identical to that used in Chapters 6 and 7. The data is manipulated using three schemes: undersampling, oversampling, and SMOTE [180]. Over- and undersampling are performed using the `ovun.sample()` function in the ROSE package [199] in R using the default settings which yield balanced classes. SMOTE sampling is done with the `SMOTE` function in the DMwR package [200] in R. The `perc.over` and `perc.under` parameters are set to 200 and 150 respectively following recommendations in [180,200] for obtaining balanced classes given the imbalance ratio of our dataset.

8.3 Methodology

We continue to use area under the curve or AUC for model selection. To determine the merits of sampling techniques we obtain sampling results for representative models from each algorithm. Essentially two different classes of datasets have been used

to make models for readmission prediction so far. The first uses a very large number of sparse variables while the second is a few aggregated variables. We used the high-dimensional data with Random Forests and support vector machines and add a generalized linear regression model (logistic regression denoted further as GLM) for completeness here.

To test how sampling schemes perform we choose representative data sets for each class of data and run these through three sampling schemes and compare the performance results to the original. The sampling schemes are over-sampling, under-sampling, and SMOTE sampling all of which were discussed in Section 2.6.

The high dimensional data for Random Forests was the max lab results. For GLM we chose variables using a simple approach. We ranked the variables according to the Random Forests' variable importance and then added variables in increasing amounts and assessed performance using 10-fold cross-validation with average error as our metric. We added variables in amounts beginning with 5 then 10×2^i where $i = 1, 2, \ldots, 9$ and then all of the variables as the final run. The optimal number of variables chosen by 10-fold cross validation was 1,280. For SVM we use the lab result indicator data set. For the low dimensional aggregated variables we use the same models aggregated variables as in Chapter 7 with the same three algorithms as with the high dimensional comparisons.

The high dimensional datasets we used were chosen because the algorithms performed very well on them in terms of AUC as reported earlier. Our purpose for assessing sampling schemes is to see if they can improve performance over current approaches. It is most helpful if sampling can improve those models that we have already found to be the best in their class. If the sampling schemes cannot improve them then we know that we can save the time and resources required to build them by

using the original data.

We build our models as in Chapters 6 and 7 using 50 tests sets split 2/3 training and 1/3 test set. We average the AUCs of the 50 runs and then compare using the Wilcoxon rank sum test. We also compare performance using ROC curves.

8.4 Results

In Table 8.1 we quickly see that for the majority of cases sampling does not improve performance. The p-values correspond to one-way ($>$) Wilcoxon signed rank test [196] of the AUCs of the highest average sampling method in the row against the sampling method with the second highest average AUC. We utilized the Holm-Bonferroni correction for multiple comparisons [107].

Among the high-dimensional data sets, only the GLM model on the original dataset did not significantly outperform the second best sampling method. In GLM's case, it was SMOTE. Under the principle of parsimony we argue that though it isn't significantly greater, it is nominally greater and requires less effort and so the original data set suffices. We conclude, then, that for high dimensional datasets across Random Forests, logistic regression, and support vector machines, sampling does not meaningfully improve performance.

For the aggregated dataset the story starts a bit differently. Both the GLM and SVM models performed nominally better than the next best methods using over-sampling. For GLM the next best method was the original data and despite being nominally better it was not significantly better. SVM however, was significantly better using over-sampling than its next best method which was SMOTE sampling. Random Forests had no distinguishable performance difference between the original and the over-sampled

Table 8.1: Comparison of select models from each algorithm across three sampling methods.

Model	Orig.	Over	Under	SMOTE	p-value
RF Max Lab	**0.8719**	0.8641	0.8357	0.8252	< 0.001
GLM Max Lab	**0.6710**	0.6587	0.6015	0.6707	0.48
SVM Lab Res.	**0.882**	0.88	0.8192	0.8252	0.044
RF Agg.	**0.865**	0.865	0.8218	0.8353	0.48
GLM Agg.	0.6674	**0.6675**	0.6658	0.6530	0.48
SVM Agg.	0.6975	**0.7074**	0.6895	0.7022	< 0.001

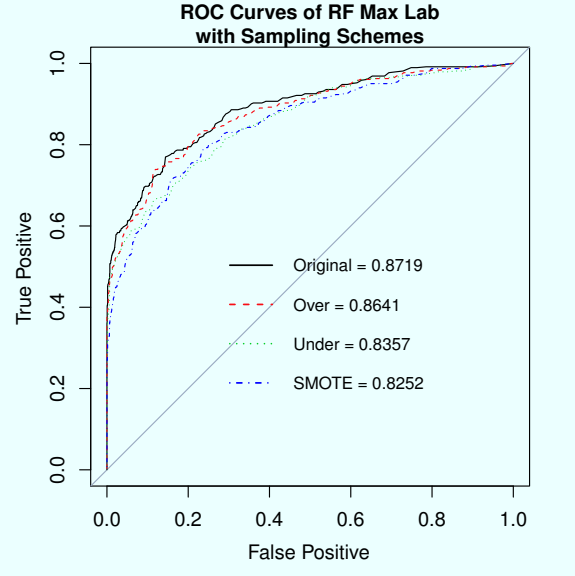

Figure 8.1: ROC curves comparing the sampling schemes on the Random Forest max lab results.

dataset. As before using the principle of economy, we choose the original dataset as it requires no effort and obtains the same results.

To visualize the results we present ROC plots for each row of Table 8.1. The ROC curves in Figure 8.1 show that generally the original dataset dominates the other sampling methods with a slight area of overlap with the over-sampling method. The ROC curves seem to reflect what the average AUCs reflect.

Figure 8.2 show severe overlap of the original dataset with the other sampling meth-

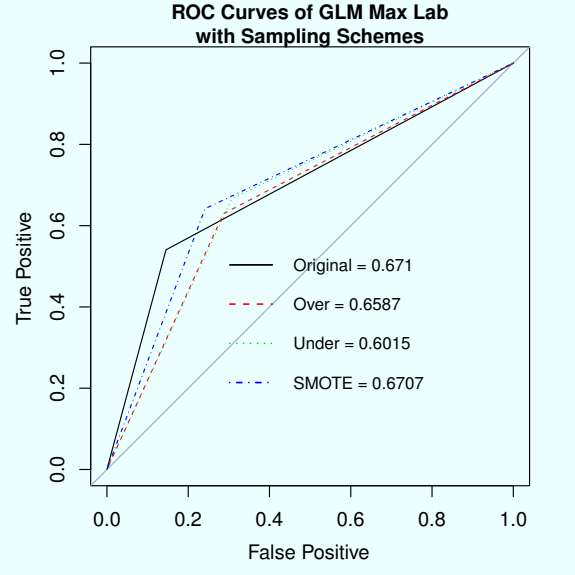

Figure 8.2: ROC curves comparing the sampling schemes on the GLM max lab results.

ods. The ROC curves are strangely straight.

Figure 8.3 shows the strong performance of SVM with all curves seeming very close to the upper left corner. The original dataset seems to dominate for most of the false positive range then overlaps with the other schemes showing that the reweighing is having some effect. Generally, the ROC curves seem to reflect what the average AUCs reflect.

Figure 8.4 is the first of the figures showing the aggregated variable results. The Random Forests results show much more overlap for the aggregated data in original form. The undersampling method seems to be clearly dominated by the others. The oversampling and original data both overlap each other in a way making their performance difficult to distinguish from one another, which was also reflected in their AUCs.

Figure 8.5 shows overlap for all the sampling methods except for SMOTE which

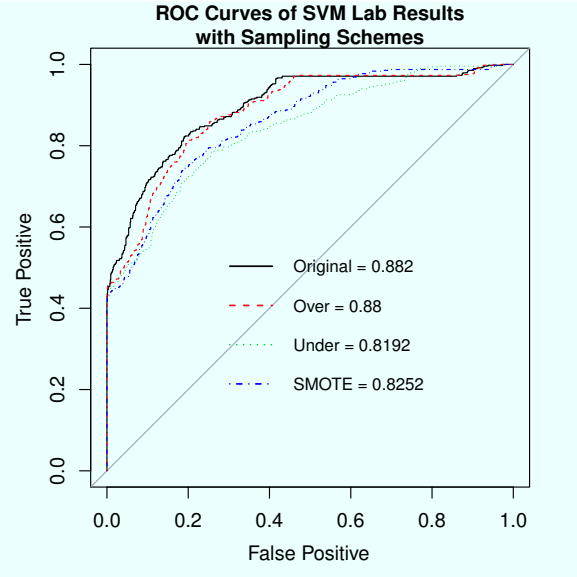

Figure 8.3: ROC curves comparing the sampling schemes on the SVM lab result indicator variables.

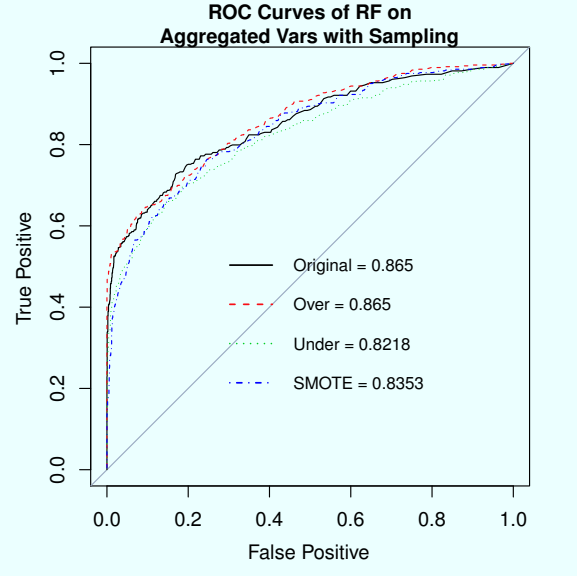

Figure 8.4: ROC curves comparing the sampling schemes on the Random Forest aggregated variables.

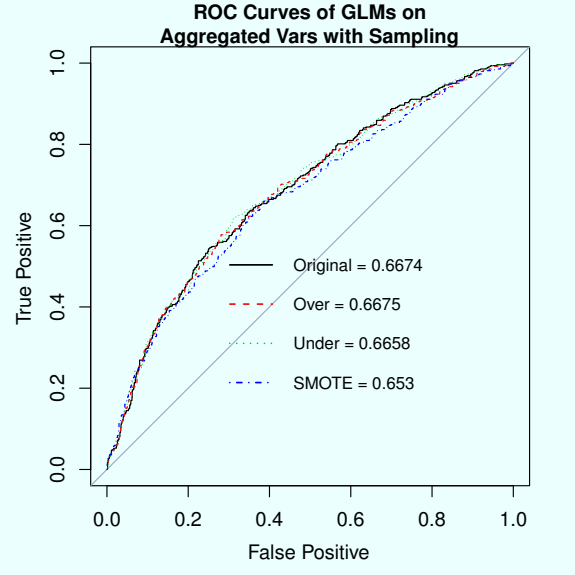

Figure 8.5: ROC curves comparing the sampling schemes on the GLM aggregated variables.

appears dominated by the others except in the lower false positive range.

Figure 8.6 shows a strong pattern for the original dataset using SVM. The original data does well until about 0.25 on the false positive range and then dips down well below all of the sampling methods. At around the same point SMOTE seems to dominate though crossing paths with oversampling. Oversampling does slightly better in the lower range than SMOTE and thus accounts for its higher AUC scores on average.

8.5 Conclusion

Generally speaking, we see that sampling may be beneficial for lower dimensional and denser datasets for SVM and GLM models. However, the Random Forests model using the original dataset vastly outperforms both the SVM and GLM models. For the low dimensional datasets, Random Forests with the original data are the model and data

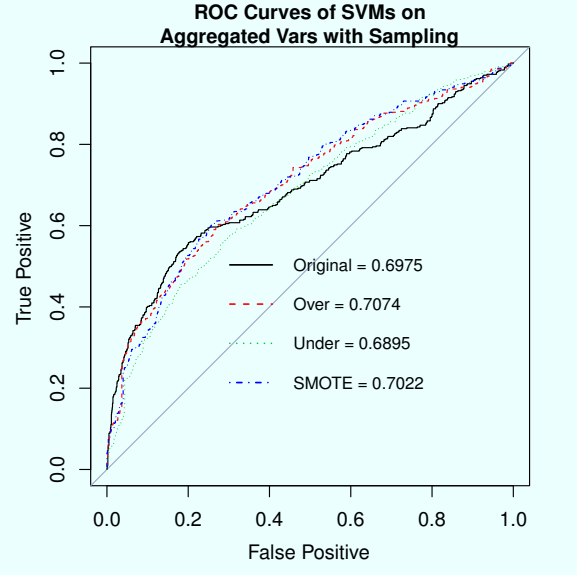

Figure 8.6: ROC curves comparing the sampling schemes on the SVM aggregated variables.

of choice. Our final conclusion then, is that sampling schemes using both high- and low-dimensional data are not necessary when using the appropriate algorithm.

Chapter 9

Contributions & Conclusion

To improve quality of care both during a visit and upon release, this dissertation has developed a methodological approach for measuring consistency in care during patients' hospital stays as well as provided a new paradigm for modeling unplanned thirty-day readmissions. The following sections document the contributions for each and discuss future work as well.

For ease of use, throughout the next sections, hyperlinks to appropriate tables and sections are provided to see results that pertain to the stated contributions when clicked.

9.1 Multidisciplinary Research Contributions

This dissertation provides multidisciplinary contributions that both shape how researchers can approach the problems discussed as well as provides specific guidance on many parts of the techniques and methods themselves. The first contribution is the diagnosis-response framework (§3.3). Administrative healthcare data are not in a clas-

sic table format and consist of many categorical variables. Other datasets with plagued by similar data structures may benefit from our framework. The diagnosis-response matrix simplifies a complicated data structure, assesses many diagnoses simultaneously, and allows for measuring care variation in a scalable way. The DRM captures key associations in highly categorical data and allows for computation of many different measures of care variation (e.g., vector space similarities, invented measures, MMD, etc.). Though the framework was applied to medical administrative data in this dissertation, applications beyond the medical realm exist. Any dataset rich in categorical data with variation or patterns are excellent candidates. This research could be extended to text mining and other data mining problems. The DRM reduces complexity while capturing variations in the system data.

Another contribution of the DRM is that it can be constructed for any level of an organization assuming the data is available (§3.4.1, 3.4.2, 3.4.3, & 3.12). This allows the within group metrics and between group comparisons performed to plausibly be performed at any organization level.

The between group variation methodology allows for a much more exploratory approach in variation research than current approaches. Whereas prior methods were too aggregated, time intensive or interrupted care, this approach allows for a look into all of the procedures used to treat a given diagnosis (§3.3). Prior methods require specifying beforehand the procedures to be explored. This approach allows researchers to keep all of the procedures in, leaving all of the proverbial doors open allowing for unknown procedures to be more deeply explored. The same would generalize to datasets where many factors must be compared at once. The methodology ranks the row variables in order of importance and still provides statistical statements of significance (E.g., Tables 3.3 & 3.6). If Random Forests is used, the within metrics may also be ranked relative

to other variables using variable importance and provide insight into their contribution toward some outcome of interest (§6.4.2). Also, using column vectors of the DRM, variation within a distribution can be measured and validated using predictive models (§6.4.1 & 6.4.2).

We also provide a variable selection technique to expedite the model creation process for RFs. Many current methods use some variant of univariate model significance as a feature selection tool. Using Random Forests we were able to retain most of the variables and have highly predictive models (§8.3 & Tables 6.5, 6.6, 6.7, 6.8 & 6.11) but in only two steps as opposed to prior methods with multiple steps.

One contribution we provide to the broader data mining community is that SVMs can be used on high-dimensional data and perform very well (Chapter 7). This contradicts earlier studies that used simulated high-dimensional array data. A related contribution is that SVMs performed well on sparse data where the relevance of the variables was largely unknown (Chapters 5, 6 & 7). In addition, we show that sampling techniques do not improve performance when used on high-dimensional data (Chapter 8). This was found to be true with SVMs, RFs, and even GLMs (Table 8.1). The only case where performance improved was for low-dimensional, high density variables SVM models (Table 8.1). But even when SVM performance improved, they did not perform as well as Random Forests. We found Random Forests to be robust to data dimensionality. They performed very well with both sparse, high-dimensional data as well as dense, low-dimensional data (Table 6.12). It was found that RFs can drastically improve the performance of GLMs on even low-dimensional datasets (Table 8.1).

Another broader contribution is that various representations attempting to capture the same idea do not all perform the same. For example, we captured the idea that extremes in lab results are bad and tend to indicate worsening physiological condition.

We captured that idea using many lab result representations including the maximum lab results value, the minimum value, the sum of squared abnormality and several others (§6.2.3). Choosing the appropriate representation had both nominally and statistically significant effects on the overall model performance (Tables 6.5, 6.10, 6.11). These effects only compounded when combined with various representations of medication administration (Tables 6.8 & 6.12). These findings emphasize the need to be creative in variable representation and not settle for the first representation that comes to mind. Related is that a few well chosen representations, combined with well-chosen algorithms can perform very well. One of the lessons that this research gives is that, often, counts may be the best option for variables representation. Our aggregate models all did very well, especially when compared to much more complicated and higher dimensional models (Table 6.14).

9.2 Health Informatics Contributions

Readmissions Contributions We were able to compare model performance for predicting thirty-day readmissions across three algorithms (Tables 7.1, 7.2, 7.3, & 8.1), more than a dozen data representations (Chapters 5, 6, & 7), and three sampling techniques (Chapter 8). Our findings move the body of work forward as no such comparison has been done to date. What little has been done with Random Forests and support vector machines had not shown great leaps in performance. We can say that sampling schemes are unhelpful (Table 8.1), and for pure predictive performance, the best algorithm is support vector machines (Table 7.3). If interpretability is important then researchers may use Random Forests with minimal loss in predictive performance (Table 7.3).

As for parameterizing Random Forests models, we have shown that 500 trees are sufficient and using the square root of the number of variables is appropriate for the number of variables to randomly select at each node (§5.3). For support vector machines we have shown that the radial basis function with $cost = 1$ and $\gamma = 0.0625$ is optimal (§7.2).

We have contributed several novel non-medical and derived clinical variables that were demonstrated to be important (§6.2.3). Included are using the last lab result from a visit, a lab result indicator variable, a lab's sum of squared abnormality, and the maximum and minimum lab results. Medication variables included all three of the variation metrics, as well as the maximum number of medications administered on a day, the number of days a medication was administered and the number of medications administered during the entire visit. These novel representations all provided significant lift over base model performance (Tables 6.6 & 6.7).

The idea of using the variation of a procedure distribution as a predictor in addition to the procedure counts is a novel contribution. We were able to validate the usefulness and predictive power of the within cohort metrics in many of the readmissions models (Tables 6.7, 6.11, & 7.2). Prior heart failure thirty-day readmissions modeling efforts used indicator variables. We replaced them with and showed that counts of variables can be more informative than mere indicators (Table 6.5). In some cases, the variation metrics were more predictive of readmission than the counts of the procedures and other metrics (Tables 6.7 & 6.11). This implies that the variation itself is informative for predictive models.

In addition to creating novel count variables and many derivative variables (§6.2.3), we were able to include all of the variables in a way that enhanced model performance (Chapters 6 & 7). In the case of SVMs we could retain all of the variables and allow

the algorithm to use what is necessary for the support vector optimization problem (Tables 7.1, 7.2, & 7.3).

This change in modeling paradigm has made a significant difference in performance and could shift the tide toward more high performance algorithms and novel uses of data. Most striking is the difference in the style of approach between this research and most prior approaches. The overriding themes of this dissertation have been simple metrics (§6.2.3) that capture a lot of information (counts, variation metrics), simple data representations that still allow for information retention (DRM plus variable representations) and powerful models capable of capitalizing on the high-dimensional data created by the DRM. Prior approaches aggregated too much through single number indices or very low dimensional models, or they used algorithms that were unable to capture nuances in higher-dimensional data.

We have shown that variables that could be proxies for a patient's conditions contribute to models' performance. For example, the degree of sickness could be indicated using length of stay, while how chronically ill they are could be shown via the number of prior visits (Table 6.13). We were able to confirm earlier findings that the Payor was an important variable (Table 6.13) and could indicate socio-economic influence on patient quality of care. We were also able to demonstrate that explicit clinical variables, in our case laboratory results, drastically improve model performance. We were also able to show that even implicit clinical variables such as medications administered also provide significant lift over base models.

The top four variables across four top models are the number of prior visits, the payor, the age and the length of stay in that order. The fifth and sixth ranks have 401.9, and 511.9 as well as 414.01 which are hypertension, unspecified pleural effusion, and coronary atherosclerosis of a native coronary artery respectively (Table 6.13). We were

able to demonstrate a surprising amount of agreement in variable importance across four well performing models. We were also able to show that control variables such as Gender and Ethnicity also play an important role in readmission prediction (Table 6.13). Lab results that appeared consistently as important variables were albumin serum, blood urea nitrogen, B-type Natriuretic Peptide, and creatine levels (Table 6.13).

Through our work, we have been able to call into question the traditional approach of modeling thirty-day readmissions. While prior work has used feature selection techniques, it has focused on a narrow set of variables to inform the model. We have expanded the feature space to include variables that may seem irrelevant if only barely informative. Given the performance of so many readmissions models, it seems clear that knowing beforehand which variables will improve performance is not straightforward; there are many factors at play. Logistic regression models are a global model and when the response surface is truly complex a global model will only be a rough approximation and suffer performance-wise. Ensemble methods take advantage of many local learners, or trees in our case, and aggregate a lot of "weak" learners in a way that reduces their variance. This approach has done well on our high-dimensional, sparse data set. We covered a large amount of modeling space, and this should give pause to the field and help researchers consider changing their modeling paradigm.

Our most obvious contributions are the models themselves. We have dozens of models all of which have an AUC of at least 0.8. More than half of those models have an AUC at least 0.85. The best model has an AUC of 0.883, more than 0.15 greater than the previous reported best model for a general heart failure population. That is a 20.8% increase in AUC. The best overall model used the support vector machine algorithm with the lab result indicator variables combined with the total number of each

medications variables all combined with the both dataset. Our best Random Forests model was 0.8722 with the max lab results combined with the total visit counts for each medicine with the both dataset. This was a more than 0.14 increase which is a 19.5% increase in AUC.

Care Variation Contributions We have shown that a more generalized method of capturing variation in treatment between patient cohorts is possible (Chapter 3) and useful (Paragraph 3.4.1). We found that measuring variation within a patient cohort is statistically significant in predicting visit charge (Table 4.4) and thirty-day readmissions (Chapters 6 & 7). Depending on the model choice, within cohort variation metrics can outperform many other candidate metrics in predicting readmissions. This further validates their use and helps us understand how variation in procedures may contribute to readmissions. The Gini metric was a strong performer (Table 6.7) which only adds to its reputation as a metric of inequality within a distribution and expands its applications.

We have shown that variation between patient cohorts attributable to principal procedures is less than that attributable to all procedures (§3.4.1 & 3.4.2). We have produced lists of procedures that are significantly different under principal procedures, all procedures, and also intersections and complements of the two (Tables 3.3, 3.6, 3.8, 3.7, 3.10, & 3.9). These provide valid starting points for exploring whether that variation has any effects on care outcomes, which treatment protocols should be reviewed for improving consistency, and a greater understanding of the treatment variation in the heart failure cohort.

We were also able to demonstrate that (when principal procedures are used) there is evidence to suggest that there is a relationship between the specificity of a diagnosis

designation (NEC/NOS vs. specific subclassifcation) and the likelihood of diagnosis cohorts found to be significantly different (Paragraph 3.4.1). In other words, we have presented evidence that those diagnoses that are less specific may in general have more variation. This was not true when all procedures were included (Paragraph 3.4.2).

We were also able to show that the most common specified and unspecified atherosclerosis diagnoses cohorts were not significantly differently treated in terms of procedure counts (§3.4.3). We then showed that when combined, those patients with Atherosclerosis were treated differently for heart failure than those without it (§3.4.3). As a validation piece, we demonstrated that physicians of different specialty could be discovered using our methods (§3.12).

For the first time, statistical comparisons across an entire treatment profile, or distribution of procedures, is possible. This provides a fuller description of the discrepancy between two patient cohorts. We have found that even after a Holm-Bonferroni adjustment, we identified nineteen diagnosis cohorts that were significantly different when only principal procedures addressing heart failure are used (§3.4.1). Thirty-one were identified as significantly different at least one time. When all procedures addressing congestive heart failure are included we identified 79 that were significantly different at least one time (§3.4.2). We identified 54 that were significantly different even after Holm-Bonferroni corrections for making 2,383 comparisons. This kind of mass identification has been heretofore unreported. This demonstrates the highly exploratory nature of this methodology and allows a tool for finding unexpected or unusual results. Traditional methods greatly reduce that possibility as they can only handle a very few diagnoses and few procedures at a time.

9.3 Future Work

The work that may stem from the findings and approaches presented in this dissertation are many. We presented many diagnosis cohorts that demonstrated significant variation in procedure application. Further exploration into the procedures that most vary between the groups and why could provide fruitful medical insight. Additionally, more work could be done to choose patient cohorts that represent important problems in the medical sphere for care variation analysis. Greater exploration of how variation within cohorts may be predictive could yield interesting results. We used within cohort variation as predictors of visit charge and thirty-day readmission, however, many other outcomes of interest exist and variation may provide key information into their prediction. We also only explored three candidate within cohort variation metrics, but there may be better metrics. We also did not explore the properties of these metrics in great detail, rather, we focused on their predictiveness. More work on how each metric assesses variation could provide insight into choosing better metrics in the future.

While we have explored a large number of modeling scenarios we have not exhausted all possibilities. Other algorithms or methods may prove to perform better, such as boosting trees. We have shown the power well-chosen, simple variables can have in improving predictive models. Surely, there are other simple variables that may be formed and further improve model performance. We have also limited the features we have used to those generally available at hospitals in the US. However, as electronic health record systems continue to unfold, we will have greater access to more physiological data. Some data that could be explored are daily weight and blood pressure readings, while more involved would be heart rate data. Incorporating these types of variables would be a logical next step. We have limited our research to congestive heart failure patients. There are many diseases that represent major issues to the US and other

countries. The robustness of the recommendations in this dissertation could be tested against these other major groups.

9.4 Finale

The highest aim of this research is to save lives by improving quality of care. Our contributions include a paradigm shift that drastically improves readmissions prediction, methods for assessing care variation across numerous diagnoses and procedures in a statistically valid way, and direct applications and demonstrations of the methods. We have shown algorithms, data transformations, and modeling approaches that are best suited for the data most hospitals in the US have available to predict thirty-day readmissions. We have provided methods to take that same data, and perform exploratory care variation analyses for any level of a care organization. This dissertation has provided means for assessing quality of care during a stay while providing models to better assure the transition after discharge occurs at the proper time.

Chapter 10

Bibliography

[1] M. A. Vedomske, D. E. Brown, and J. H. Harrison Jr., "Random forests on ubiquitous data for heart failure 30-day readmissions prediction," in *2013 12th International Conference on Machine Learning and Applications (ICMLA)*, vol. (being indexed), (Miami, FL), IEEE, Dec. 2013.

[2] P. Plsek, "Redesigning health care with insights from the science of complex adaptive systems," *Crossing the Quality Chasm: A New Heath System for the 21st Century: National Academy of Sciences*, p. 309–322, 2000.

[3] W. H. Organization, "Quality of care: a process for making strategic choices in health systems," 2006.

[4] C. for Medicare & Medicaid Services, *Roadmap for quality measurement in the traditional Medicare fee-for-service program.* 2009.

[5] M. Heron, "Deaths: leading causes for 2007.," *National vital statistics reports: from the Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System*, vol. 59, no. 8, p. 1, 2011.

[6] W. G. MEMBERS, D. Lloyd-Jones, R. Adams, M. Carnethon, G. De Simone, T. B. Ferguson, K. Flegal, E. Ford, K. Furie, A. Go, K. Greenlund, N. Haase, S. Hailpern, M. Ho, V. Howard, B. Kissela, S. Kittner, D. Lackland, L. Lisabeth, A. Marelli, M. McDermott, J. Meigs, D. Mozaffarian, G. Nichol, C. O'Donnell, V. Roger, W. Rosamond, R. Sacco, P. Sorlie, R. Stafford, J. Steinberger, T. Thom, S. Wasserthiel-Smoller, N. Wong, J. Wylie-Rosett, Y. Hong, for the American Heart Association Statistics Committee, and S. S. Subcommittee, "Heart disease and stroke statistics—2009 update," *Circulation*, vol. 119, pp. e21 –e181, Jan. 2009.

[7] S. S. D. US Census Bureau, "US census bureau population clocks," Oct. 2011. US Census Bureau US and World Population Clocks.

[8] A. H. Association, "American heart association financial commitiment," July 2012.

[9] T. D. A. of Health Care, "PERCENT OF PATIENTS READMITTED WITHIN 30 DAYS OF DISCHARGE, BY COHORT," Online Report Generation 1, The Dartmouth Institute for Health Policy and Clinical Practice, 2010.

[10] U. Public Health Sciences, "UVa clinical data repository," May 2012.

[11] J. R. Schubart and J. S. Einbinder, "Evaluation of a data warehouse in an academic health sciences center," *International Journal of Medical Informatics*, vol. 60, pp. 319–333, Dec. 2000.

[12] N. C. f. H. S. (US), C. o. C. Classifications, C. o. Professional, H. Activities, and W. H. Organization, *The International Classification of Diseases, 9th Revision, Clinical Modification: ICD. 9. CM.* Commission on Professional and Hospital Activities, 1978.

[13] M. Beebe, J. A. Dalton, M. Espronceda, D. D. Evans, R. L. Glenn, and G. Green, *CPT: Standard: Current Procedural Terminology.* Amer Medical Assn, 2007.

[14] DHHS, "Guidance regarding methods for de-identification of protected health information in accordance with the health insurance portability and accountability act (HIPAA) privacy rule," 2013.

[15] D. E. Beck and D. A. Margolin, "Physician coding and reimbursement," *The Ochsner Journal*, vol. 7, no. 1, pp. 8–15, 2007. PMID: 21603473 PMCID: PMC3096340.

[16] S. Levant and C. DeFrances, "Electronic collection of inpatient and ambulatory hospital care data: national hospital care survey," in *Proceedings of the 13th Annual International Conference on Digital Government Research*, dg.o '12, (New York, NY, USA), p. 200–205, ACM, 2012.

[17] A. Sboner and C. F. Aliferis, "Modeling clinical judgment and implicit guideline compliance in the diagnosis of melanomas using machine learning," in *AMIA Annual Symposium Proceedings*, vol. 2005, p. 664, 2005.

[18] D. Eisenhower, N. A. Mathiowetz, and D. Morganstein, "Recall error: Sources and bias reduction techniques," in *Measurement Errors in Surveys* (P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A.thiowetz, and S. Sudman, eds.), p. 125–144, John Wiley & Sons, Inc., 2004.

[19] C. Mettlin, G. Murphy, M. Cunningham, and H. Menck, "The national cancer data base report on race, age, and region variations in prostate cancer treatment," *Cancer*, vol. 80, no. 7, p. 1261–1266, 1997.

[20] Z. Irwin, M. Arthur, R. Mullins, and R. Hart, "Variations in injury patterns, treatment, and outcome for spinal fracture and paralysis in adult versus geriatric

patients," *Spine*, vol. 29, no. 7, pp. 796–802, 2004.

[21] A. Majeed, K. Moser, and R. Maxwell, "Age, sex and practice variations in the use of statins in general practice in england and wales," *Journal of Public Health*, vol. 22, no. 3, pp. 275–279, 2000.

[22] M. Komajda, F. Follath, K. Swedberg, J. Cleland, J. Aguilar, A. Cohen-Solal, R. Dietz, A. Gavazzi, W. Van Gilst, R. Hobbs, *et al.*, "The euroheart failure survey programme—a survey on the quality of care among patients with heart failure in europe," *European Heart Journal*, vol. 24, no. 5, pp. 464–474, 2003.

[23] K. Fox, S. Goodman, W. Klein, D. Brieger, P. Steg, O. Dabbous, and A. Avezum, "Management of acute coronary syndromes. variations in practice and outcome. findings from the global registry of acute coronary events (grace)," *European heart journal*, vol. 23, no. 15, pp. 1177–1189, 2002.

[24] L. Bosco, B. Gerstman, and D. Tomita, "Variations in the use of medication for the treatment of childhood asthma in the michigan medicaid population, 1980 to 1986.," *Chest*, vol. 104, no. 6, pp. 1727–1732, 1993.

[25] K. McGuire, J. Harrast, H. Herkowitz, and J. Weinstein, "Geographic variation in the surgical treatment of degenerative cervical disc disease: American board of orthopedic surgery quality improvement initiative; part ii candidates," *Spine*, vol. 37, no. 1, pp. 57–66, 2012.

[26] R. Ghandour, M. Kogan, S. Blumberg, J. Jones, and J. Perrin, "Mental health conditions among school-aged children: geographic and sociodemographic patterns in prevalence and treatment," *Journal of Developmental & Behavioral Pediatrics*, vol. 33, no. 1, pp. 42–54, 2012.

[27] J. Mindell, N. Shelton, M. Roth, M. Chaudhury, and E. Falaschetti, "Persistent regional variation in treatment of hypertension," *Public Health*, vol. 126, pp. 317–323, 2012.

[28] G. O'Connor, H. Quinton, N. Traven, L. Ramunno, T. Dodds, T. Marciniak, and J. Wennberg, "Geographic variation in the treatment of acute myocardial infarction," *JAMA: the journal of the American Medical Association*, vol. 281, no. 7, pp. 627–633, 1999.

[29] L. Pilote, R. Califf, S. Sapp, D. Miller, D. Mark, W. Weaver, J. Gore, P. Armstrong, E. Ohman, and E. Topol, "Regional variation across the united states in the management of acute myocardial infarction," *New England journal of medicine*, vol. 333, no. 9, pp. 565–572, 1995.

[30] J. Wennberg, "Unwarranted variations in healthcare delivery: implications for academic medical centres," *Bmj*, vol. 325, no. 7370, pp. 961–964, 2002.

[31] W. Johnson and H. Jacobe, "Morphea in adults and children cohort ii: Patients with morphea experience delay in diagnosis and large variation in treatment," *Journal of the American Academy of Dermatology*, 2012.

[32] J. Tu, D. Ko, H. Guo, J. Richards, N. Walton, M. Natarajan, H. Wijeysundera, D. So, D. Latter, C. Feindel, *et al.*, "Determinants of variations in coronary revascularization practices," *Canadian Medical Association Journal*, vol. 184, no. 2, pp. 179–186, 2012.

[33] G. Fonarow, C. Yancy, N. Albert, A. Curtis, W. Stough, M. Gheorghiade, J. Heywood, M. McBride, M. Mehra, C. O'Connor, *et al.*, "Heart failure care in the outpatient cardiology practice setting," *Circulation: Heart Failure*, vol. 1, no. 2, pp. 98–106, 2008.

[34] S. Stewart, C. Demers, D. Murdoch, K. McIntyre, M. MacLeod, S. Kendrick, S. Capewell, and J. McMurray, "Substantial between-hospital variation in outcome following first emergency admission for heart failure," *European heart journal*, vol. 23, no. 8, pp. 650–657, 2002.

[35] P. P. Goodney, L. L. Travis, D. Malenka, K. K. Bronner, F. L. Lucas, J. L. Cronenwett, D. C. Goodman, and E. S. Fisher, "Regional variation in carotid artery stenting and endarterectomy in the medicare population," *Circulation: Cardiovascular Quality and Outcomes*, vol. 3, pp. 15–24, Jan. 2010.

[36] R. Giugliano, C. Camargo Jr, D. Lloyd-Jones, J. Zagrodsky, J. Alexis, K. Eagle, V. Fuster, and C. O'Donnell, "Elderly patients receive less aggressive medical and invasive management of unstable angina: potential impact of practice guidelines," *Archives of internal medicine*, vol. 158, no. 10, pp. 1113–1120, 1998.

[37] T. L. Schreiber, A. Elkhatib, C. L. Grines, and W. W. O'Neill, "Cardiologist versus internist management of patients with unstable angina: Treatment patterns and outcomes," *Journal of the American College of Cardiology*, vol. 26, pp. 577–582, Sept. 1995.

[38] A. Potosky, S. Saxman, R. Wallace, and C. Lynch, "Population variations in the initial treatment of non–small-cell lung cancer," *Journal of clinical oncology*, vol. 22, no. 16, pp. 3261–3268, 2004.

[39] H. Gold and A. Dick, "Variations in treatment for ductal carcinoma in situ in elderly women," *Medical care*, vol. 42, no. 3, pp. 267–275, 2004.

[40] N. Hébert-Croteau, J. Brisson, J. Latreille, C. Blanchette, and L. Deschênes, "Variations in the treatment of early-stage breast cancer in quebec between 1988

and 1994," *Canadian Medical Association Journal*, vol. 161, no. 8, pp. 951–955, 1999.

[41] S. Rathore, A. Berger, K. Weinfurt, M. Feinleib, W. Oetgen, B. Gersh, and K. Schulman, "Race, sex, poverty, and the medical treatment of acute myocardial infarction in the elderly," *Circulation*, vol. 102, no. 6, pp. 642–648, 2000.

[42] K. Huybrechts, K. Rothman, M. Brookhart, R. Silliman, S. Crystal, T. Gerhard, and S. Schneeweiss, "Variation in antipsychotic treatment choice across us nursing homes," *Journal of Clinical Psychopharmacology*, vol. 32, no. 1, pp. 11–17, 2012.

[43] S. Wheeler, W. Carpenter, J. Peppercorn, A. Schenck, M. Weinberger, and A. Biddle, "Structural/organizational characteristics of health services partly explain racial variation in timeliness of radiation therapy among elderly breast cancer patients," *Breast Cancer Research and Treatment*, pp. 1–13, 2012.

[44] R. Stafford, D. Saglam, and D. Blumenthal, "National patterns of angiotensin-converting enzyme inhibitor use in congestive heart failure," *Archives of internal medicine*, vol. 157, no. 21, pp. 2460–2464, 1997.

[45] J. Whittle, J. Conigliaro, C. Good, and R. Lofgren, "Racial differences in the use of invasive cardiovascular procedures in the department of veterans affairs medical system," *New England Journal of Medicine*, vol. 329, no. 9, pp. 621–627, 1993.

[46] P. Jong, Y. Gong, P. Liu, P. Austin, D. Lee, and J. Tu, "Care and outcomes of patients newly hospitalized for heart failure in the community treated by cardiologists compared with other specialists," *Circulation*, vol. 108, no. 2, pp. 184–191, 2003.

[47] V. Nambudiri, M. Landrum, E. Lamont, B. McNeil, S. Bozeman, S. Freedland, and N. Keating, "Understanding variation in primary prostate cancer treatment within the veterans health administration," *Urology*, vol. 76, pp. 537–545, 2012.

[48] M. McCarthy, R. Ding, J. Pines, C. Terwiesch, M. Sattarian, J. Hilton, J. Lee, and S. Zeger, "Provider variation in fast track treatment time," *Medical Care*, vol. 50, no. 1, pp. 43–49, 2012.

[49] S. Jaglal, P. Sherry, D. Chua, and J. Schatzker, "Temporal trends and geographic variations in surgical treatment of femoral neck fractures," *The Journal of trauma*, vol. 43, no. 3, pp. 475–479, 1997.

[50] A. Simons, R. Ker, S. Groshen, C. Gee, G. Anthone, A. Ortega, P. Vukasin, R. Ross, and R. Beart, "Variations in treatment of rectal cancer," *Diseases of the colon & rectum*, vol. 40, no. 6, pp. 641–646, 1997.

[51] D. Winchester, H. Menck, and D. Winchester, "National treatment trends for ductal carcinoma in situ of the breast," *Archives of Surgery*, vol. 132, no. 6, pp. 660–665, 1997.

[52] B. Boari, E. Mari, M. Gallerani, F. Fabbian, M. Pala, R. Tiseo, and R. Manfredini, "Temporal variation of heart failure hospitalization: does it exist?," *Reviews in cardiovascular medicine*, vol. 12, no. 4, pp. 211–218, 2011.

[53] M. Gallerani, B. Boari, F. Manfredini, and R. Manfredini, "Seasonal variation in heart failure hospitalization," *Clinical cardiology*, vol. 34, pp. 389–394, 2011.

[54] S. E. Reis, R. Holubkov, D. Edmundowicz, D. M. McNamara, K. A. Zell, K. M. Detre, and A. M. Feldman, "Treatment of patients admitted to the hospital with congestive heart failure: Specialty-related disparities in practice patterns and

outcomes," *Journal of the American College of Cardiology*, vol. 30, pp. 733–738, July 1997.

[55] C. Green, J. Wheeler, and F. LaPorte, "Clinical decision making in pain management: Contributions of physician and patient characteristics to variations in practice," *The Journal of Pain*, vol. 4, no. 1, pp. 29–39, 2003.

[56] M. Edep, E. Martin, M. Shah, B. Nihir, M. Tateo, M. Ida, M. Massie, and M. Barry, "Differences between primary care physicians and cardiologists in management of congestive heart failure: relation to practice guidelines," *Journal of the American College of Cardiology*, vol. 30, no. 2, pp. 518–526, 1997.

[57] D. Baker, R. Hayes, B. Massie, and C. Craig, "Variations in family physicians' and cardiologists' care for patients with heart failure," *American heart journal*, vol. 138, no. 5, pp. 826–834, 1999.

[58] F. Peters-Klimm, G. Laux, S. Campbell, T. Müller-Tasch, N. Lossnitzer, J. Schultz, A. Remppis, J. Jünger, and C. Nikendei, "Physician and patient predictors of evidence-based prescribing in heart failure: A multilevel study," *PloS one*, vol. 7, no. 2, p. e31082, 2012.

[59] F. Lucas, B. Sirovich, P. Gallagher, A. Siewers, and D. Wennberg, "Variation in cardiologists' propensity to test and treat," *Circulation: Cardiovascular Quality and Outcomes*, vol. 3, no. 3, pp. 253–260, 2010.

[60] A. Fuat, A. Hungin, and J. Murphy, "Barriers to accurate diagnosis and effective management of heart failure in primary care: qualitative study," *Bmj*, vol. 326, no. 7382, pp. 196–201, 2003.

[61] S. S. Coughlin, "Recall bias in epidemiologic studies," *Journal of Clinical Epidemiology*, vol. 43, no. 1, pp. 87–91, 1990.

[62] V. Nambudiri, M. Landrum, E. Lamont, B. McNeil, S. Bozeman, S. Freedland, and N. Keating, "Understanding variation in primary prostate cancer treatment within the veterans health administration," *Urology*, vol. 76, p. 537–545, 2012.

[63] T. J. Smith, L. Penberthy, C. E. Desch, M. Whittemore, C. Newschaffer, B. E. Hillner, D. McClish, and S. M. Retchin, "Differences in initial treatment patterns and outcomes of lung cancer in the elderly," *Lung Cancer*, vol. 13, pp. 235–252, Dec. 1995.

[64] W. H. Hall, A. B. Jani, J. K. Ryu, S. Narayan, and S. Vijayakumar, "The impact of age and comorbidity on survival outcomes and treatment patterns in prostate cancer," *Prostate Cancer Prostatic Dis*, vol. 8, pp. 22–30, Feb. 2005.

[65] G. C. Fonarow, J. T. Heywood, P. A. Heidenreich, M. Lopatin, and C. W. Yancy, "Temporal trends in clinical characteristics, treatments, and outcomes for heart failure hospitalizations, 2002 to 2004: findings from acute decompensated heart failure national registry (ADHERE)," *American Heart Journal*, vol. 153, pp. 1021–1028, June 2007.

[66] W. T. Abraham, G. C. Fonarow, N. M. Albert, W. G. Stough, M. Gheorghiade, B. H. Greenberg, C. M. O'Connor, J. L. Sun, C. W. Yancy, and J. B. Young, "Predictors of in-hospital mortality in patients hospitalized for heart failure: Insights from the organized program to initiate lifesaving treatment in hospitalized patients with heart failure (OPTIMIZE-HF)," *Journal of the American College of Cardiology*, vol. 52, pp. 347–356, July 2008.

[67] M. G. Bourassa, O. Gurné, S. I. Bangdiwala, J. K. Ghali, J. B. Young, M. Rousseau, D. E. Johnstone, and S. Yusuf, "Natural history and patterns of current practice in heart failure," *Journal of the American College of Cardiology*,

vol. 22, pp. A14–A19, Oct. 1993.

[68] T. D. Wang, C. Plaisant, A. J. Quinn, R. Stanchak, S. Murphy, and B. Shneiderman, "Aligning temporal data by sentinel events: discovering patterns in electronic health records," in *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, CHI '08, (New York, NY, USA), p. 457–466, ACM, 2008.

[69] C. Plaisant, S. Lam, B. Shneiderman, M. S. Smith, D. Roseman, G. Marchand, M. Gillam, C. Feied, J. Handler, and H. Rappaport, "Searching electronic health records for temporal patterns in patient histories: A case study with microsoft amalga," *AMIA Annual Symposium Proceedings*, vol. 2008, pp. 601–605, 2008. PMID: 18999158 PMCID: 2655947.

[70] S. Mani, W. Shankle, M. Dick, and M. Pazzani, "Two-stage machine learning model for guideline development," *Artificial Intelligence in Medicine*, 1998.

[71] S. Mani, W. R. Shankle, M. B. Dick, and M. J. Pazzani, "Two-stage machine learning model for guideline development," *Artificial Intelligence in Medicine*, vol. 16, no. 1, pp. 51–71, 1999.

[72] D. Riano, "Time-independent rule-based guideline induction," in *ECAI*, vol. 16, p. 535, 2004.

[73] M. Toussi, J. Lamy, P. Le Toumelin, and A. Venot, "Using data mining techniques to explore physicians' therapeutic decisions when clinical guidelines do not provide recommendations: methods and example for type 2 diabetes," *BMC medical informatics and decision making*, vol. 9, no. 1, p. 28, 2009.

[74] M.-J. Huang, M.-Y. Chen, and S.-C. Lee, "Integrating data mining with case-based reasoning for chronic diseases prognosis and diagnosis," *Expert Systems*

with Applications, vol. 32, pp. 856–867, Apr. 2007.

[75] S. Patil and Y. Kumaraswamy, "Extraction of significant patterns from heart disease warehouses for heart attack prediction," *IJCSNS International Journal of Computer Science and Network Security*, vol. 9, no. 2, p. 228–235, 2009.

[76] Q. R. Huang, Z. Qin, S. Zhang, and C. M. Chow, "Clinical patterns of obstructive sleep apnea and its comorbid conditions: A data mining approach," *Journal of Clinical Sleep Medicine : JCSM : Official Publication of the American Academy of Sleep Medicine*, vol. 4, pp. 543–550, Dec. 2008. PMID: 19110883 PMCID: 2603531.

[77] J. Li, A. W.-c. Fu, H. He, J. Chen, H. Jin, D. McAullay, G. Williams, R. Sparks, and C. Kelman, "Mining risk patterns in medical data," in *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, KDD '05, (New York, NY, USA), p. 770–775, ACM, 2005.

[78] C. J. Ryan, H. A. DeVon, R. Horne, K. B. King, K. Milner, D. K. Moser, J. R. Quinn, A. Rosenfeld, S. Y. Hwang, and J. J. Zerwic, "Symptom clusters in acute myocardial infarction," *Nursing Research*, vol. 56, pp. 72–81, Mar. 2007.

[79] A. McCutcheon, *Latent class analysis*. No. 07-064 in Sage University Paper series on Quantitative Applications in the Social Sciences, United States of America: Sage, 1987.

[80] S. Ting, W. Wang, Y. Tse, and W. Ip, "Knowledge elicitation approach in enhancing tacit knowledge sharing," *Industrial Management & Data Systems*, vol. 111, pp. 1039–1064, Aug. 2011.

[81] K. Pearson, "X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably

supposed to have arisen from random sampling," *Philosophical Magazine Series 5*, vol. 50, no. 302, pp. 157–175, 1900.

[82] W. G. Cochran, "The chi-squared test of goodness of fit," *The Annals of Mathematical Statistics*, vol. 23, pp. 315–345, Sept. 1952.

[83] R. A. Fisher, "On the mathematical foundations of theoretical statistics," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 222, pp. 309–368, 1922. ArticleType: primary_article / Full publication date: 1922 / Copyright © 1922 The Royal Society.

[84] I. Campbell, "Chi-squared and Fisher–Irwin tests of two-by-two tables with small sample recommendations," *Statistics in medicine*, vol. 26, no. 19, p. 3661–3675, 2007.

[85] W. G. Cochran, "SOME METHODS FOR STRENGTHENING THE COMMON x2 TESTS," *Biometrics*, vol. 10, pp. 417–451, 1954.

[86] F. Yates, "Contingency tables involving small numbers and the chi-squared test," *Supplement to the Journal of the Royal Statistical Society*, vol. 1, pp. 217–235, Jan. 1934.

[87] N. Mantel, "Chi-square tests with one degree of freedom; extensions of the mantel-haenszel procedure," *Journal of the American Statistical Association*, vol. 58, no. 303, p. 690–700, 1963.

[88] J. T. Roscoe and J. A. Byars, "An investigation of the restraints with respect to sample size commonly imposed on the use of the chi-square statistic," *Journal of the American Statistical Association*, vol. 66, no. 336, p. 755–759, 1971.

[89] J. H. Steiger, A. Shapiro, and M. W. Browne, "On the multivariate asymptotic distribution of sequential chi-square statistics," *Psychometrika*, vol. 50, no. 3, p. 253–263, 1985.

[90] N. MANTEL and J. L. FLEISS, "Minimum expected cell size requirements for the mantel-haenszel one-degree-of-freedom chi-square test and a related rapid procedure," *American Journal of Epidemiology*, vol. 112, no. 1, p. 129–134, 1980.

[91] M. Dwass, "Modified randomization tests for nonparametric hypotheses," *The Annals of Mathematical Statistics*, vol. 28, pp. 181–187, Mar. 1957. ArticleType: research-article / Full publication date: Mar., 1957 / Copyright © 1957 Institute of Mathematical Statistics.

[92] G. A. Barnard, "Discussion on professor bartlett's paper," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 25, p. 294, Jan. 1963.

[93] M. S. Bartlett, "The spectral analysis of point processes," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 25, pp. 264–296, Jan. 1963. ArticleType: research-article / Full publication date: 1963 / Copyright © 1963 Royal Statistical Society.

[94] A. C. A. Hope, "A simplified monte carlo significance test procedure," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 30, pp. 582–598, Jan. 1968. ArticleType: research-article / Full publication date: 1968 / Copyright © 1968 Royal Statistical Society.

[95] P. Hall and D. M. Titterington, "The effect of simulation order on level accuracy and power of monte carlo tests," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 51, pp. 459–467, Jan. 1989.

[96] B. V. North, D. Curtis, and P. C. Sham, "A note on the calculation of empirical p values from monte carlo procedures," *American journal of human genetics*, vol. 71, no. 2, p. 439, 2002.

[97] W. J. Ewens, "On estimating p values by monte carlo methods," *American Journal of Human Genetics*, vol. 72, pp. 496–498, Feb. 2003. PMID: 12596794 PMCID: PMC529334.

[98] J. Besag and P. J. Diggle, "Simple monte carlo tests for spatial pattern," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 26, pp. 327–333, Jan. 1977.

[99] D. A. Roff and P. Bentzen, "The statistical analysis of mitochondrial DNA polymorphisms: chi 2 and the problem of small samples.," *Molecular Biology and Evolution*, vol. 6, pp. 539–545, Sept. 1989. PMID: 2677600.

[100] L. A. Waller, D. Smith, J. E. Childs, and L. A. Real, "Monte carlo assessments of goodness-of-fit for ecological simulation models," *Ecological Modelling*, vol. 164, pp. 49–63, June 2003.

[101] G. K. Sandve, E. Ferkingstad, and S. Nygård, "Sequential monte carlo multiple testing," *Bioinformatics*, vol. 27, pp. 3235–3241, Dec. 2011. PMID: 21998154.

[102] J. Besag and P. Clifford, "Sequential monte carlo p-values," *Biometrika*, vol. 78, pp. 301–304, June 1991.

[103] J. Besag and P. Clifford, "Generalized monte carlo significance tests," *Biometrika*, vol. 76, pp. 633–642, Dec. 1989.

[104] J. Friedman, "On multivariate goodness-of-fit and two-sample testing," *Proceedings of Phystat2003, http://www. slac. stanford. edu/econf C*, vol. 30908, 2004.

[105] K.-H. Jockel, "Finite sample properties and asymptotic efficiency of monte carlo tests," *The Annals of Statistics*, vol. 14, pp. 336–347, Mar. 1986.

[106] R. D. C. Team, "R," 2011.

[107] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, vol. 6, pp. 65–70, Jan. 1979.

[108] Y. Hochberg, "A sharper bonferroni procedure for multiple tests of significance," *Biometrika*, vol. 75, no. 4, p. 800–802, 1988.

[109] T. V. Perneger, "What's wrong with bonferroni adjustments," *BMJ*, vol. 316, pp. 1236–1238, Apr. 1998. PMID: 9553006.

[110] K. J. Rothman, "No adjustments are needed for multiple comparisons," *Epidemiology*, vol. 1, pp. 43–46, Jan. 1990.

[111] G. G. Koch and S. A. Gansky, "Statistical considerations for multiplicity in confirmatory protocols," *Drug Information Journal*, vol. 30, no. 2, p. 523–534, 1996.

[112] A. J. Sankoh, M. F. Huque, and S. D. Dubey, "Some comments on frequently used multiple endpoint adjustment methods in clinical trials," *Statistics in medicine*, vol. 16, no. 22, p. 2529–2542, 1997.

[113] M. A. Proschan and M. A. Waclawiw, "Practical guidelines for multiplicity adjustment in clinical trials," *Controlled clinical trials*, vol. 21, no. 6, p. 527–539, 2000.

[114] R. Bender and S. Lange, "Adjusting for multiple testing—when and how?," *Journal of Clinical Epidemiology*, vol. 54, pp. 343–349, Apr. 2001.

[115] C. Gini, "Measurement of inequality of incomes," *The Economic Journal*, vol. 31, no. 121, p. 124–126, 1921.

[116] J. L. Gastwirth, "The estimation of the lorenz curve and gini index," *The Review of Economics and Statistics*, vol. 54, pp. 306–316, Aug. 1972. ArticleType: research-article / Full publication date: Aug., 1972 / Copyright © 1972 The MIT Press.

[117] P. Protection and A. C. Act, "Patient protection and affordable care act," *Public Law*, no. 111-148, 2010.

[118] Vaduganathan M, Bonow RO, and Gheorghiade M, "Thirty-day readmissions: The clock is ticking," *JAMA*, vol. 309, pp. 345–346, Jan. 2013.

[119] R. D. Kociol, L. Liang, A. F. Hernandez, L. H. Curtis, P. A. Heidenreich, C. W. Yancy, G. C. Fonarow, and E. D. Peterson, "Are we targeting the right metric for heart failure? comparison of hospital 30-day readmission rates and total episode of care inpatient days," *American Heart Journal*, vol. 165, pp. 987–994.e1, June 2013.

[120] K. E. Joynt and A. K. Jha, "Thirty-day readmissions — truth and consequences," *New England Journal of Medicine*, vol. 366, no. 15, pp. 1366–1369, 2012. PMID: 22455752.

[121] A. M. Epstein, "Revisiting readmissions, changing the incentives for shared accountability," *New England Journal of Medicine*, vol. 360, no. 14, pp. 1457–1459, 2009. PMID: 19339727.

[122] M. R. Chassin, J. M. Loeb, S. P. Schmaltz, and R. M. Wachter, "Accountability measures, using measurement to promote quality improvement," *New England Journal of Medicine*, vol. 363, no. 7, pp. 683–688, 2010. PMID: 20573915.

[123] S. F. Jencks, M. V. Williams, and E. A. Coleman, "Rehospitalizations among patients in the medicare fee-for-service program," *New England Journal of Medicine*, vol. 360, no. 14, pp. 1418–1428, 2009. PMID: 19339721.

[124] L. Horwitz, C. Partovian, Z. Lin, J. Herrin, J. Grady, M. Conover, J. Montague, C. Dillaway, K. Bartczak, and J. Ross, "Hospital-wide (all-condition) 30-day risk-standardized readmission measure," tech. rep., Centers for Medicare & Medicaid Services, 2011.

[125] Fonarow GC, Adams KF, Abraham WT, Yancy CW, Boscardin W, ADHERE Scientific Advisory Committee, Study Group, and and Investigators f, "Risk stratification for in-hospital mortality in acutely decompensated heart failure: Classification and regression tree analysis," *JAMA*, vol. 293, pp. 572–580, Feb. 2005.

[126] P. S. Keenan, S.-L. T. Normand, Z. Lin, E. E. Drye, K. R. Bhat, J. S. Ross, J. D. Schuur, B. D. Stauffer, S. M. Bernheim, A. J. Epstein, Y. Wang, J. Herrin, J. Chen, J. J. Federer, J. A. Mattera, Y. Wang, and H. M. Krumholz, "An administrative claims measure suitable for profiling hospital performance on the basis of 30-day all-cause readmission rates among patients with heart failure," *Circulation: Cardiovascular Quality and Outcomes*, vol. 1, pp. 29–37, Sept. 2008. PMID: 20031785.

[127] R. Amarasingham, B. J. Moore, Y. P. Tabak, M. H. Drazner, C. A. Clark, S. Zhang, W. G. Reed, T. S. Swanson, Y. Ma, and E. A. Halm, "An automated model to identify heart failure patients at risk for 30-day readmission or death using electronic medical record data:," *Medical Care*, vol. 48, pp. 981–988, Nov. 2010.

[128] B. G. Hammill, L. H. Curtis, G. C. Fonarow, P. A. Heidenreich, C. W. Yancy, E. D. Peterson, and A. F. Hernandez, "Incremental value of clinical data beyond claims data in predicting 30-day outcomes after heart failure hospitalization," *Circulation: Cardiovascular Quality and Outcomes*, vol. 4, pp. 60–67, Jan. 2011. PMID: 21139093.

[129] D. S. Lee, P. C. Austin, J. L. Rouleau, P. P. Liu, D. Naimark, and J. V. Tu, "Predicting mortality among patients hospitalized for heart failure," *JAMA: The Journal of the American Medical Association*, vol. 290, pp. 2581 –2587, Nov. 2003.

[130] A. G. Au, F. A. McAlister, J. A. Bakal, J. Ezekowitz, P. Kaul, and C. van Walraven, "Predicting the risk of unplanned readmission or death within 30 days of discharge after a heart failure hospitalization," *American Heart Journal*, vol. 164, pp. 365–372, Sept. 2012.

[131] E. Bradley, O. Yakusheva, L. Horwitz, H. Sipsma, and J. Fletcher, "Identifying patients at increased risk for unplanned readmission," *Medical Care*, vol. 51, pp. 761–766, Sept. 2013.

[132] M. Rothman, S. Rothman, and J. Beals IV, "Development and validation of a continuous measure of patient condition using the electronic medical record," *Journal of Biomedical Informatics*, vol. In Press, 2013.

[133] C. M. DesRoches, C. Worzala, M. S. Joshi, P. D. Kralovec, and A. K. Jha, "Small, nonteaching, and rural hospitals continue to be slow in adopting electronic health record systems," *Health Affairs*, vol. 31, pp. 1092–1099, May 2012. PMID: 22535503.

[134] M. Zuniga, D. Anderson, and K. Alexander, "Heart disease and stroke in rural america," *Rural healthy people*, 2010.

[135] C. D. McNaughton, S. P. Collins, S. Kripalani, R. Rothman, W. H. Self, C. Jenkins, K. Miller, P. Arbogast, A. Naftilan, and R. S. Dittus, "Low numeracy is associated with increased odds of 30-day emergency department or hospital recidivism for patients with acute heart failure," *Circulation: Heart Failure*, vol. 6, no. 1, p. 40–46, 2013.

[136] A. F. Hernandez, M. A. Greiner, G. C. Fonarow, B. G. Hammill, P. A. Heidenreich, C. W. Yancy, E. D. Peterson, and L. H. Curtis, "Relationship between early physician follow-up and 30-day readmission among medicare beneficiaries hospitalized for heart failure," *JAMA: The Journal of the American Medical Association*, vol. 303, no. 17, p. 1716–1722, 2010.

[137] W. C. Levy, D. Mozaffarian, D. T. Linker, S. C. Sutradhar, S. D. Anker, A. B. Cropp, I. Anand, A. Maggioni, P. Burton, M. D. Sullivan, B. Pitt, P. A. Poole-Wilson, D. L. Mann, and M. Packer, "The seattle heart failure model prediction of survival in heart failure," *Circulation*, vol. 113, pp. 1424–1433, Mar. 2006. PMID: 16534009.

[138] D. Logeart, G. Thabut, P. Jourdain, C. Chavelas, P. Beyne, F. Beauvais, E. Bouvier, and A. C. Solal, "Predischarge b-type natriuretic peptide assay for identifying patients at high risk of re-admission after decompensated heart failure," *Journal of the American College of Cardiology*, vol. 43, pp. 635–641, Feb. 2004.

[139] H. M. Krumholz, Y.-T. Chen, Y. Wang, V. Vaccarino, M. J. Radford, and R. I. Horwitz, "Predictors of readmission among elderly survivors of admission with heart failure," *American heart journal*, vol. 139, no. 1, p. 72–77, 2000.

[140] H. M. Krumholz, J. Amatruda, G. L. Smith, J. A. Mattera, S. A. Roumanis, M. J. Radford, P. Crombie, and V. Vaccarino, "Randomized trial of an education

and support intervention to preventreadmission of patients with heart failure," *Journal of the American College of Cardiology*, vol. 39, no. 1, p. 83–89, 2002.

[141] M. D. Naylor, D. A. Brooten, R. L. Campbell, G. Maislin, K. M. McCauley, and J. S. Schwartz, "Transitional care of older adults hospitalized with heart failure: a randomized, controlled trial," *Journal of the American Geriatrics Society*, vol. 52, no. 5, p. 675–684, 2004.

[142] M. W. Rich, V. Beckham, C. Wittenberg, C. L. Leven, K. E. Freedland, and R. M. Carney, "A multidisciplinary intervention to prevent the readmission of elderly patients with congestive heart failure," *New England Journal of Medicine*, vol. 333, no. 18, p. 1190–1195, 1995.

[143] S. Stewart, J. E. Marley, and J. D. Horowitz, "Effects of a multidisciplinary, home-based intervention on planned readmissions and survival among patients with chronic congestive heart failure: a randomised controlled study," *The Lancet*, vol. 354, no. 9184, p. 1077–1083, 1999.

[144] E. Hsich, E. Z. Gorodeski, E. H. Blackstone, H. Ishwaran, and M. S. Lauer, "Identifying important risk factors for survival in patient with systolic heart failure using random survival forests," *Circulation: Cardiovascular Quality and Outcomes*, vol. 4, pp. 39–45, Jan. 2011. PMID: 21098782.

[145] Y.-J. Son, H.-G. Kim, E.-H. Kim, S. Choi, and S.-K. Lee, "Application of support vector machine for prediction of medication adherence in heart failure patients," *Healthcare Informatics Research*, vol. 16, no. 4, p. 253, 2010.

[146] S. Cholleti, A. Post, J. Gao, X. Lin, W. Bornstein, D. Cantrell, and J. Saltz, "Leveraging derived data elements in data analytic models for understanding

and predicting hospital readmissions," in *AMIA Annual Symposium Proceedings*, vol. 2012, p. 103, 2012.

[147] X. Lin, *A Predictive Random Forest Model on Hospital 30-Day Readmission using Electronic Health Records*. MSPH, Emory University, May 2012.

[148] Kansagara D, Englander H, Salanitro A, and et al, "Risk prediction models for hospital readmission: A systematic review," *JAMA*, vol. 306, pp. 1688–1698, Oct. 2011.

[149] D. Powers, "The problem of area under the curve," in *2012 International Conference on Information Science and Technology (ICIST)*, pp. 567–573, 2012.

[150] T. Hastie, R. Tibshirani, and J. Friedman, "13nsupervised learning," in *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*, Springer, 2nd ed. 2009. corr. 3rd printing 5th printing. ed., Feb. 2009.

[151] V. López, A. Fernández, S. García, V. Palade, and F. Herrera, "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics," *Information Sciences*, vol. 250, pp. 113–141, Nov. 2013.

[152] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, pp. 1145–1159, July 1997.

[153] H. M. Krumholz, A. R. Merrill, E. M. Schone, G. C. Schreiner, J. Chen, E. H. Bradley, Y. Wang, Y. Wang, Z. Lin, B. M. Straube, M. T. Rapp, S.-L. T. Normand, and E. E. Drye, "Patterns of hospital performance in acute myocardial infarction and heart failure 30-day mortality and readmission," *Circulation:*

Cardiovascular Quality and Outcomes, vol. 2, pp. 407–413, Sept. 2009. PMID: 20031870.

[154] E. F. Philbin and T. G. DiSalvo, "Prediction of hospital readmission for heart failure: development of a simple risk score based on administrative data," *Journal of the American College of Cardiology*, vol. 33, pp. 1560–1566, May 1999.

[155] N. Meadem, N. Verbiest, K. Zolfaghar, J. Agarwal, S.-C. Chin, and S. B. Roy, "Exploring preprocessing techniques for prediction of risk of readmission for congestive heart failure patients," in *Data Mining and Healthcare Workshop, in conjunction with the 19th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2013.

[156] R. Gildersleeve and P. Cooper, "Development of an automated, real time surveillance tool for predicting readmissions at a community hospital," *Methods Inf Med*, vol. 46, no. 5, p. 553–557, 2007.

[157] K. Zolfaghar, N. Meadem, A. Teredesai, S. Roy, S.-C. Chin, and B. Muckian, "Big data solutions for predicting risk-of-readmission for congestive heart failure patients," in *2013 IEEE International Conference on Big Data*, pp. 64–71, Oct. 2013.

[158] R. Agrawal and R. Srikant, "Mining sequential patterns," in *Data Engineering, 1995. Proceedings of the Eleventh International Conference on*, p. 3–14, 1995.

[159] L. Breiman and A. Cutler, "Random forests-classification description," *Department of Statistics Homepage*, 2007.

[160] R. Genuer, J.-M. Poggi, and C. Tuleau-Malot, "Variable selection using random forests," *Pattern Recognition Letters*, vol. 31, pp. 2225–2236, Oct. 2010.

[161] A. Liaw and M. Wiener, "Classification and regression by randomForest," *R News*, vol. 2, pp. 18–22, Dec. 2002.

[162] U. Grömping, "Variable importance assessment in regression: Linear regression versus random forest," *The American Statistician*, vol. 63, no. 4, pp. 308–319, 2009.

[163] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures: Illustrations, sources and a solution," *BMC Bioinformatics*, vol. 8, p. 25, Jan. 2007. PMID: 17254353.

[164] C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, "Conditional variable importance for random forests," *BMC Bioinformatics*, vol. 9, p. 307, July 2008. PMID: 18620558.

[165] V. Vapnik, *The nature of statistical learning theory*. springer, 1996.

[166] H. He and E. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, pp. 1263–1284, Sept. 2009.

[167] M. J. Pazzani, C. J. Merz, P. M. Murphy, K. Ali, T. Hume, and C. Brunk, "Reducing misclassification costs.," in *ICML*, vol. 94, p. 217–225, 1994.

[168] P. Domingos, "Metacost: A general method for making classifiers cost-sensitive," in *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, p. 155–164, ACM, 1999.

[169] B. C. Wallace and I. J. Dahabreh, "Improving class probability estimates for imbalanced data," *Knowledge and Information Systems*, pp. 1–20.

[170] Y. Tang, Y.-Q. Zhang, N. Chawla, and S. Krasser, "SVMs modeling for highly imbalanced classification," *IEEE Transactions on Systems, Man, and Cybernet-*

ics, Part B: Cybernetics, vol. 39, no. 1, pp. 281–288, 2009.

[171] Y. F. Roumani, J. H. May, D. P. Strum, and L. G. Vargas, "Classifying highly imbalanced ICU data," *Health Care Management Science*, vol. 16, pp. 119–128, June 2013.

[172] N. V. Chawla, N. Japkowicz, and A. Kotcz, "Editorial: Special issue on learning from imbalanced data sets," *SIGKDD Explor. Newsl.*, vol. 6, p. 1–6, June 2004.

[173] N. Japkowicz, "Learning from imbalanced data sets: a comparison of various strategies," in *AAAI workshop on learning from imbalanced data sets*, vol. 68, Menlo Park, CA, 2000.

[174] M. Khalilia, S. Chakraborty, and M. Popescu, "Predicting disease risks from highly imbalanced data using random forest," *BMC Medical Informatics and Decision Making*, vol. 11, p. 51, July 2011. PMID: 21801360.

[175] X.-Y. Liu, J. Wu, and Z.-H. Zhou, "Exploratory undersampling for class-imbalance learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 39, pp. 539–550, Apr. 2009.

[176] P. Cao, J. Yang, W. Li, D. Zhao, and O. Zaiane, "Ensemble-based hybrid probabilistic sampling for imbalanced data learning in lung nodule CAD," *Computerized Medical Imaging and Graphics*, vol. 38, pp. 137–150, Apr. 2014.

[177] J.-J. Liao, C.-H. Shih, T.-F. Chen, and M.-F. Hsu, "An ensemble-based model for two-class imbalanced financial problem," *Economic Modelling*, vol. 37, pp. 175–183, Feb. 2014.

[178] M. Galar, A. Fernández, E. Barrenechea, and F. Herrera, "EUSBoost: enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling,"

Pattern Recognition, vol. 46, pp. 3460–3471, Dec. 2013.

[179] X. Jiang, R. El-Kareh, and L. Ohno-Machado, "Improving predictions in imbalanced data using pairwise expanded logistic regression," *AMIA Annual Symposium Proceedings*, vol. 2011, pp. 625–634, 2011. PMID: 22195118 PMCID: PMC3243279.

[180] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, p. 321–357, 2002.

[181] R. Blagus and L. Lusa, "SMOTE for high-dimensional class-imbalanced data," *BMC Bioinformatics*, vol. 14, p. 106, Mar. 2013. PMID: 23522326.

[182] D. M. Eddy, "Variations in physician practice: the role of uncertainty," *Health Affairs*, vol. 3, pp. 74–89, May 1984. PMID: 6469198.

[183] J. Wennberg and A. Gittelsohn, "Small area variations in health care delivery a population-based health information system can guide planning and regulatory decision-making," *Science*, vol. 182, no. 4117, p. 1102–1108, 1973.

[184] K. Scully, "Clinical data repository help," June 2012.

[185] M. G. Haviland, "Yates's correction for continuity and the analysis of 2 x 2 contingency tables," *Statistics in Medicine*, vol. 9, pp. 363–367, Apr. 1990.

[186] T. K. Landauer, P. W. Foltz, and D. Laham, "Introduction to latent semantic analysis," in *Discourse Processes*, vol. 25, pp. 259–284, 1998.

[187] D. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," *The Journal of Machine Learning Research*, vol. 3, p. 993–1022, 2003.

[188] H. Analytics, "Healthcare information and management systems society analytics annual study," 2013.

[189] J. Lindenfeld, N. Albert, J. Boehmer, S. Collins, J. Ezekowitz, M. Givertz, S. Katz, M. Klapholz, D. Moser, J. Rogers, *et al.*, "HFSA 2010 comprehensive heart failure practice guideline.," *Journal of cardiac failure*, vol. 16, no. 6, p. e1, 2010.

[190] A. Elixhauser and C. Steiner, "Readmissions to U.S. hospitals by dignosis, 2010," Statistical Brief 153, Agency for Healthcare Research and Quality, Rockville, MD, Apr. 2013.

[191] M. E. Charlson, P. Pompei, K. L. Ales, and C. R. MacKenzie, "A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation* 1," *Journal of chronic diseases*, vol. 40, no. 5, p. 373–383, 1987.

[192] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, p. 5–32, 2001.

[193] C. Chen, A. Liaw, and L. Breiman, "Using random forest to learn imbalanced data," *University of California, Berkeley*, 2004.

[194] K. J. O'Malley, K. F. Cook, M. D. Price, K. R. Wildes, J. F. Hurdle, and C. M. Ashton, "Measuring diagnoses: ICD code accuracy," *Health Services Research*, vol. 40, no. 5p2, p. 1620–1639, 2005.

[195] E. P. McCarthy, L. I. Iezzoni, R. B. Davis, R. H. Palmer, M. Cahalane, M. B. Hamel, K. Mukamal, R. S. Phillips, and D. T. Davies, "Does clinical evidence support ICD-9-CM diagnosis coding of complications?," *Medical Care*, vol. 38, pp. 868–876, Aug. 2000. ArticleType: research-article / Full publication date: Aug., 2000 / Copyright © 2000 Lippincott Williams & Wilkins.

[196] M. Hollander, D. A. Wolfe, and E. Chicken, *Nonparametric Statistical Methods*. John Wiley & Sons, Nov. 2013.

[197] B. Wallace, K. Small, C. Brodley, and T. Trikalinos, "Class imbalance, redux," in *2011 IEEE 11th International Conference on Data Mining (ICDM)*, pp. 754–763, 2011.

[198] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch, "e1071: Misc functions of the department of statistics (e1071)." 2012.

[199] N. Lunardon, G. Menardi, N. Torelli, M. N. Lunardon, and M. Suggests, "Package 'ROSE'," 2013.

[200] L. Torgo and M. L. Torgo, "Package 'DMwR'," 2013.

[201] J. Ayanian and E. Guadagnoli, "Variations in breast cancer treatment by patient and provider characteristics," *Breast cancer research and treatment*, vol. 40, no. 1, pp. 65–74, 1996.

[202] B. Xhyheri and R. Bugiardini, "Diagnosis and treatment of heart disease: are women different from men?," *Progress in cardiovascular diseases*, vol. 53, no. 3, pp. 227–236, 2010.