BRADBERRY, CALEB, Ph.D. A Design Science Framework for Research in Health Analytics. (2016)
Directed by Dr. Rahul Singh. 187 pp.

Data analytics provide the ability to systematically identify patterns and insights from a variety of data as organizations pursue improvements in their processes, products, and services. Analytics can be classified based on their ability to: explore, explain, predict, and prescribe. When applied to the field of healthcare, analytics presents a new frontier for business intelligence. In 2013 alone, the Centers for Medicare and Medicaid Services (CMS) reported that the national health expenditure was $2.9 trillion, representing 17.4% of the total United States GDP. The Patient Protection and Affordable Care Act of 2010 (ACA) requires all hospitals to implement electronic medical record (EMR) technologies by year 2014 (*Patient Protection and Affordable Care Act*, 2010). Moreover, the ACA makes healthcare process and outcomes more transparent by making related data readily available for research. Enterprising organizations are employing analytics and analytical techniques to find patterns in healthcare data (I. R. Bardhan & Thouin, 2013; Hansen, Miron-Shatz, Lau, & Paton, 2014). The goal is to assess the cost and quality of care and identify opportunities for improvement for organizations as well as the healthcare system as a whole. Yet, there remains a need for research to systematically understand, explain, and predict the sources and impacts of the widely observed variance in the cost and quality of care available. This is a driving motivation for research in healthcare.

This dissertation conducts a design theoretic examination of the application of advanced data analytics in healthcare. Heart Failure is the number one cause of death and the biggest contributor healthcare costs in the United States. An exploratory examination of the application of predictive analytics is conducted in order to understand the cost and quality of care provided to heart failure patients. The specific research question is addressed: *How can we improve and expand upon our understanding of the variances in the cost of care and the quality of care for heart failure?* Using state level data from the State Health Plan of North Carolina, a standard readmission model was assessed as a baseline measure for prediction, and advanced analytics were compared to this baseline. This dissertation demonstrates that advanced analytics can improve readmission predictions as well as expand understanding of the profile of a patient readmitted for heart failure. Implications are assessed for academics and practitioners.

A DESIGN SCIENCE FRAMEWORK FOR RESEARCH IN HEALTH ANALYTICS

by

Caleb Bradberry

A Dissertation Submitted to
the Faculty of The Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Greensboro
2016

Approved by

Rahul Singh
Committee Chair

To Susan Bradberry, who encouraged me to ask questions when I was young, may her memory live on through my actions.

APPROVAL PAGE

This dissertation written by Caleb Bradberry has been approved by the following Committee of the Faculty of The Graduate School at The University of North Carolina at Greensboro.

Committee Chair	Rahul Singh
Committee Members	Joyendu Bhadury
	Larry Taube
	Kathy White Loyd

June 24, 2016
Date of Acceptance by Committee

June 24, 2016
Date of Final Oral Examination

ACKNOWLEDGEMENTS

This research is made possible from the help and advice of my entire dissertation committee. I am very thankful for the patience that the committee provided throughout this entire process. I would like to thank Dr. Rahul Singh for helping to shape the research into a successful project; I would like to thank Dr. Joy Bhadury for his guidance with the mathematical aspects of this research; I would like to thank Dr. Larry Taube for his guidance in clarifying ideas and core concepts to this research; I would like to thank Dr. Kathy White-Loyd for her guidance in shaping a successful research project that will benefit both academe and industry.

I would like to extend a special thank you to the analytics team at the North Carolina State Health Plan. Their help and knowledge provide invaluable to the success of this research: Beverly Harris, Rupa Maitra, and Rajendra Urukadle all provided critical information around the industry practices, nomenclature, and processes.

Lastly, I want to acknowledge close friends and family who kept me going and provided a new set of eyes when mine were tired. Erica Martin, for not only putting up with my ramblings, but also helping me through this research; Alan Pack, for providing a fresh set of eyes.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER I

INTRODUCTION

Research Purpose

The Agency for Healthcare Research and Quality (AHRQ) provides standard benchmarks for what constitutes quality of healthcare. These benchmarks are based on four quality indicator modules (Farquhar, 2008): prevention quality indicators (PQIs), inpatient quality indicators (IQIs), patient safety indicators (PSIs), and pediatric quality indicators (PDIs). As part of the Patient Protection and Affordable Care Act (ACA) of 2010, this legislation creates motivation for care providers to make data openly available as part of the ACA's effort to improve the quality of care and reduce the cost of care. In doing so, the ACA creates an opportunity for researchers to identify new utility in healthcare data. This dissertation takes advantage of this opportunity.

To accomplish this, this dissertation utilizes researcher and practitioner efforts from three different academic and practitioner domains. From the public domain, this dissertation adopts the measurements of quality and cost of healthcare and definitions of these measurements as analytics. From the field of information systems, work that has been conducted around the notion of data analytics as it relates to providing timely and relevant insight for managers, as well as the applications of specific advanced analytics for certain scenarios for predicting and explaining. From the fields of computer science

and statistics, this dissertation adopts the established efforts of work that has been conducted to power these advanced analytics.

This dissertation examines the intersection of these three fields. In so doing, this dissertation applies the underlying machine learning algorithms using guidance from information systems, as advanced analytics to better understand variances in cost and quality of healthcare. This dissertation provides a process to inform the application of advanced analytics to healthcare administrative data.

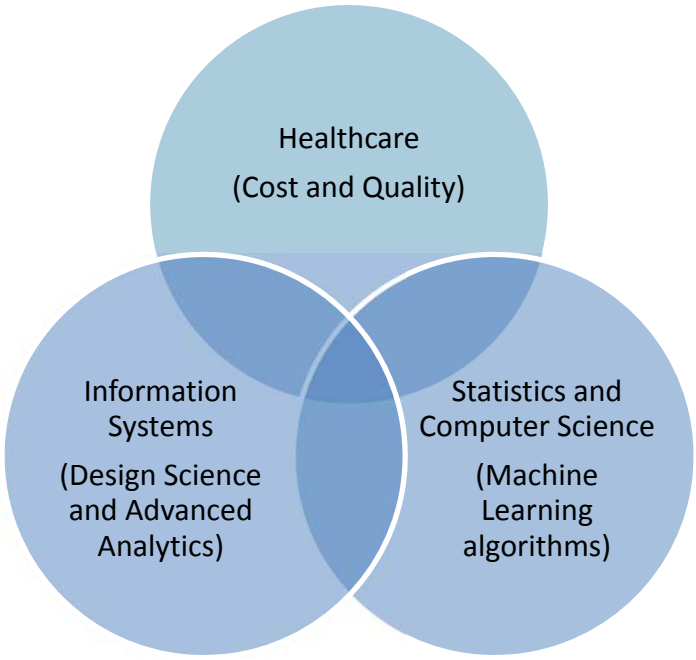


Figure 1. Dissertation Research Model.

The purpose of conducting this dissertation is to provide a class of solutions to address the class of problems that occur when examining cost and quality of healthcare.

This new knowledge benefits to academicians, practitioners, and healthcare providers. This dissertation contributes process to the body of knowledge.

Research Motivation

Data analytics afford organizations the ability to methodically assess a variety of data in the pursuit of insight around the organization's products and services as well as the end-user's experience. Analytics can be classified based on their function to: explore, explain, predict, and prescribe. These classes of analytics are employed across various dimensions of product and user.

The field of healthcare analytics presents a new frontier for data analytics. The Patient Protection and Affordable Care Act of 2010 (ACA) requires that all hospitals have implemented electronic medical record (EMR) technologies in the year 2014 (*Patient Protection and Affordable Care Act*, 2010). The ACA making data more transparent and more readily available, both researchers as well as enterprising organizations are employing analytics and analytical techniques to find patterns in healthcare data (I. R. Bardhan & Thouin, 2013; Hansen et al., 2014). The underlying motivation to find these patterns is readily apparent when simply considering the amount that the United States government spends on their public healthcare policy. In 2013 alone, the Centers for Medicare and Medicaid Services (CMS) reported that Medicare spending grew 3.4% to $585.7 billion dollars, Medicaid spending grew 6.1% to $449.4 billion, comprising of 35% of total national health expenditures (CMS, 2015a). Beyond public spending, CMS reported that hospital expenditures grew by 4.3% to $936.9 billion

and physician and clinical services grew by 3.8% to $586.7 billion. CMS reports that the national health expenditure totaled $2.9 trillion (17.4%) of total GDP for the United States in the year 2013. In the wake of this spending, if a healthcare organization can identify even a 0.1% reduction in cost, it results in a considerable amount of savings to the organization. This fact motivates this dissertation.

In healthcare literature, one of the most prominent themes when assessing the business aspect of the organization is dual objectives of cost and quality (Berwick, Nolan, & Whittington, 2008; Ma, 1994; The White House, 2013; Weisbrod, 1991). The cost of healthcare is ultimately defined by the total economic expenditure incurred by an individual, group, or organization for treatment. Data around this expenditure has been made available in part from provisions in the ACA as well as a function of CMS. The goal of the ACA in making this data available is increased transparency, but this also presents opportunity to researchers to understand the various factors contributing to cost and quality of care. This available data is grouped by in-patient, out-patient, emergency department, and pharmacy claims. Whereas the cost of healthcare is a more concrete metric, multiple avenues of research as well as practitioners provide a federally mandated standard of measures for defining quality of healthcare.

> Quality measures are tools that help us measure or quantify healthcare processes, outcomes, patient perceptions, and organizational structure and/or systems that are associated with the ability to provide high-quality health care and/or that relate to one or more quality goals for health care. These goals include: effective, safe, efficient, patient-centered, equitable, and timely care. (CMS, 2015b)

Quality indicators consist of a description, numerator, and denominator. The result of a quality indicator is a percentage based on the specification of what is being assessed. The following is an example of one of these quality of care measures adopted by AHRQ.

Measurement Description: *Heart failure patients discharged home with written instructions or educational material given to patients or caregiver at discharge or during the hospital stay addressing all of the following: activity level, diet discharge medications, follow-up appointment, weight monitoring, and what to do if symptoms worsen.*

Measurement numerator: *Number of heart failure patients with documentation that they or their caregivers were given written discharge instructions or other educational material addressing all of the following: 1. Activity level, 2. Diet, 3. Discharge, 4. Follow-up appointment, 5. Weight monitoring, 6. What to do if the symptoms worsen.*

Measurement denominator: *heart failure patients discharged home.*

From this, a quality measurement is provided: number of heart failure patients discharged from a hospital and provided information as to help improve their recover divided by the number of heart failure patients, as a ratio to determine an aspect of quality. Assuming 100 patients were provided information, and 1000 patients were discharged, the quality indicator becomes 100/1000, or 10% of the patients. Effectively, a quality indicator becomes an analytic measurement for a specific aspect of healthcare.

By viewing quality measures as analytics, a bridge is created between health care cost and quality measures and the information systems (IS) domain of analytics. IS broadly defines analytics under four umbrellas: to explain and to predict (Chen, Chiang,

& Storey, 2012; Shmueli & Koppius, 2011), as well as to explore and to prescribe (Davenport, 2013; Lavalle, Lesser, Shockley, Hopkins, & Kruschwitz, 2011; Phillips-wren, Iyer, Kulkarni, & Ariyachandra, 2015). It is of particular interest that Chen et al's work is built upon the theoretical work proposed by Gregor, in which she identifies different types of theories (Gregor, 2006) and then Chen et al. apply this theoretical basis to the creation and utilization of analytics. In so doing, analytics become rooted in theory.

While IS has identified the managerial utility behind analytics, the fields of computer science and statistics have provided established algorithms that underlie advanced analytics. This dissertation adopts these established algorithms to improve and expand upon current knowledge of specific healthcare quality metrics

This dissertation integrates the fields as follows: AHRQ provides measurements to create constructs and models to assess the efficacy of various healthcare efforts in prevention, utilization, and cost. An instance of one of these measurements is adopted, the heart failure readmission rate. Using this instance, a design theoretic process of applying advanced analytics to healthcare data is provided, demonstrated, and discussed. In so doing, this dissertation improves and expands on current understanding in variances in cost and quality of healthcare.

Research Questions

The purpose of this dissertation is to explore the utility and value of various types of advanced analytics to improve and expand understanding of variance in cost and quality of healthcare provided. This process is exploratory in nature. The process first

must define the relevant metrics and constructs within the problem domain, healthcare. Next the process identifies the relevant tools and techniques in the solution domain, advanced analytics. The specific problem domain that has been adopted is the measurement of heart failure readmission patients.

To explore this problem, data has been obtained from a health insurance plan provider. The data being used includes a unique identifier so that the history of a patient's inpatient admissions can be analyzed. Additionally, this data includes diagnosis codes, claim lines, demographic information, primary payer information, and information related to the Medicare. With this data, patients with heart failure and their hospital admissions can be assessed. With a design theoretic approach, a class of problems becomes addressed through the research questions. The research questions for this dissertation are as follows.

Research Question 1: How can we improve and expand upon our understanding of variances in the cost of care for specific health conditions?

Research Question 2: How can we improve and expand upon our understanding of variances in the quality of care for specific health conditions?

Instantiated Research Question: How can we improve and expand upon our understanding of variances in 30 day all cause readmissions of patients with heart failure?

Approach

To improve and expand upon understanding of variances in cost and quality of specific health conditions, this dissertation adopts the following approach. First, a review of design science is conducted and a research framework is synthesized to guide the process of exploration. This framework provides the necessary structure to identify models, methods, and instances for addressing classes of problems. Next, a review of applicable advanced analytics and their purpose is conducted. This review of the applicable analytics is situated in the framework. This review provides the necessary basis for establishing a boundary of the solution domain.

Next, a review of the models, methods, and instances in the problem domain is conducted. Specifically, heart failure readmission is examined in terms of the effect it has on both cost of healthcare, as well as quality of healthcare. A review of the quality indicators provided by AHRQ is provided. With readmission cost and quality established, the adopted model for predicting heart failure readmissions is examined, and then assessed.

To assess the CMS heart failure readmission model, a discussion of the data and the process by which the data was loaded is provided. The CMS model is then assessed, and found to provide analogous results. This assessment provides a necessary baseline by which advanced analytics can be assessed. To compare the CMS model to advanced analytics, a discussion of the receiver operating characteristic curve is provided. Four advanced analytics are applied to the same dataset as the CMS model, and their predictive ability is assessed using the receiver operator characteristic curve. In so doing,

improvements upon existing understand are demonstrated. To expand upon current understanding of readmission, clustering as an advanced analytic is applied to the dataset. Using the cluster memberships, the four advanced analytics are employed again, using the model the clustering results provided. In so doing, clustering is demonstrated to both improve and expand understanding.

Scope of Research

The scope of this research is to improve and expand upon understanding of variances in cost and quality for heart failure readmission patients. Specifically, 30 day, all cause heart failure readmissions are examined. Heart failure readmissions are the most costly condition for a hospital to treat in terms of cost, as well as are a quality measure of care provided as described by AHRQ. Assessing heart failure readmissions examines both cost and quality. 30 day heart failure readmissions are penalized by CMS. By instantiating the problem of predicting patients at risk of a 30 day heart failure readmission, a class of problems within the heart failure domain is also being assessed by understanding the comorbidities that manifest with readmissions.

This dissertation is scoped as follows: (1) creation of a design theoretic framework; (2) identification of the constructs, models, methods, available within the within the solution domain as they relate to prediction; (3) identification of the constructs, models, methods, and instances available in the problem domain; (4) establishment of a baseline from which this dissertation can improve and expand; (5) demonstrating improvement from the baseline; (6) demonstrating expansion from the baseline.

Dissertation Organization

This dissertation is organized as follows. Chapter 2 presents a literature review of advanced analytics and design science, the solution domain. This chapter provides a design framework, and builds a design theoretic classification of analytics based on their outcomes. Chapter 3 presents a literature review from the healthcare, quality measure, and heart failure readmission, or the problem domain. Chapter 4 describes the extraction, transformation, and loading of the healthcare administrative data, as well as replicates the CMS adopted model for predicting heart failure readmissions. Chapter 5 presents the results of four advanced analytics as an alternative to the CMS adopted model in relation to improving understanding of variances in readmission. Chapter 6 presents the results of using clustering as a mechanism to expand understanding of variances in readmission. Chapter 7 provides discussion to the objectives of Chapter 5 and 6.

CHAPTER II

A DESIGN SCIENCE FRAMEWORK FOR HEALTH ANALYTICS

This chapter presents a review of literature in the topics of advanced analytics and algorithms, and design science. The review of design science establishes a research approach that guides the identification of characteristics, design, and requirements. In addition, this chapter builds a summary of established algorithms that underlie advanced analytics based on their characteristics. This summary allows a designer to implement advanced analytics based on their requirements and capabilities, such as explain, predict, associate, etc.

The purpose of this chapter is to establish a design theoretic approach that drives this research. This review of the literature is synthesized into a research approach that frames a process to understand of the utility of applying advanced analytics to healthcare data to improve and expand existing methods of understanding cost and quality of care.

The structure of this chapter is as follows. First, seminal work to guide design science is reviewed and core concepts are adapted into this dissertation's research approach. From this review, a synthesized research approach is framed. To utilize this research approach, a mapping is presented of the characteristics of advanced analytics which examines their underlying design and requirements to position their utility as an improvement or expansion on the existing ways in which healthcare data is examined.

Design Science as a Framework

Gregor and Hevner's 2013 research essay calls for more design science efforts in Information Systems (Gregor & Hevner, 2013). They include in this call a breakdown of the types of knowledge contributions that have emerged from their review of the literature of design science. These types of knowledge contributions include: invention, improvement, exaptation, and routine design. They describe invention as new solutions for new problems, improvement as new solutions for known problems, exaptation as known solutions extended to new problems, and routine design as known solutions for known problems. Their review found that the vast majority of information systems research within the design science stream focused on improvement. The work contained in this dissertation falls into their definition of exaptation. This dissertation is taking known solutions and applying them in a new problem domain vis-à-vis advanced analytics and healthcare cost and quality.

The 1992 paper by Walls, Widmeyer, and Sawy discusses design science as a process and a product. They put forward a formulation for the application of design theories in the information systems space (Walls, Widmeyer, & Sawy, 1992). They present design theories such that while science is concerned with analysis, design is oriented towards synthesis. They compare the design process to the scientific method in that design, like theory, is a set of hypothesis that can be ultimately proven by the construction of the artificial (Simon, 1969). Walls et al. go on to discuss the differences between investigating in the natural sciences and social sciences

> The purpose is not to achieve those goals. The purpose of design theory is to support the achievement of goals.(Walls et al., 1992)

In this effort to support the achievement of goals as a viable research route, Walls et al. present seven characteristics of design science research. Table 1 defines those characteristics and presents this dissertation's implementation of the characteristics.

Table 1. Implementation of Design Science Guidelines.

Characteristic	Implementation
Design theories must deal with goals as contingencies.	If we want to [visualize and associate clusters], we need to utilize [k-means and Voronoi].
A design theory can never involve pure explanation or prediction.	This design theory explains how the process of applying machine learning can be applied; it predicts that it will improve and expand upon existing methods.
Design theories are prescriptive.	Using machine learning will provide an alternative perspective to phenomenon in healthcare data.
Design theories are composite theories which encompass kernel theories from natural science, social science, and mathematics.	Medical theories, six chronic conditions, machine learning algorithms.
Design theories answer questions of 'how to/because' as opposed to 'what is', 'what will be', and 'what should be.'	*How to* improve and expand understanding of variances in healthcare data.
Design theories show how explanatory, predictive, or normative theories can be put into practical use.	For example: a healthcare theory says that chronic heart failure will have x-number of comorbidities; this dissertation takes that knowledge and shows how that knowledge can be applied to understand cost and quality of care.
Design theories are theories of Simon's procedural rationality.	Procedures synthesized from: Walls, Widmeyer, and Sawy 1992; March and Smith, 1995; Hevner et al., 2004

The idea of design is breaks design into both a noun, and design as a verb; for design research, there is a product or an artifact or some tangible outcome. To get to this outcome, there has to be a process put in place to create the artifact. A design theoretic research must have both. The process guides the design to get to the product. The product functions as a tool for understanding if the process was followed is it is where the

outcome of the research is externally validated based on its utility. For reference, table 2 examines the components of design science to generate a product, and how this dissertation incorporates those components. Table 3 examines the components of design science as a process, the components, and the implementation of those components for this dissertation.

Table 2. Design as a Product.

COMPONENT	DEFINITION	IMPLEMENTATION
Meta-requirements	Describes the class of goals.	Class of goals: [explaining] and [predicting] the [variance of costs] and [quality] on a per metric basis
Meta-design	Describes a class of artifacts.	Algorithms that will improve and expand upon our understand of variances in cost and quality of healthcare
Kernel theories	Theories governing design *requirements*.	Literature on: heart failure, k-means clustering, explanatory analytics
Testable design product hypotheses	Tests whether meta-requirements are satisfied.	Example: k-means clustering will improve current understanding of variance in the cost of care for heart failure

Table 3. Design as a Process.

COMPONENT	DEFINITION	IMPLEMENTATION
Design Method	Procedures for artifact construction.	The process of mapping characteristics of analytics to healthcare.
Kernel theories	Theories governing the design process.	Heart failure, decision tree inductive learning, explanatory analytics

March and Smith built on this notion of design as a product and design as a process to develop a framework to position design science in the theoretical space (March & Smith, 1995a). In the original framework, design science is broken into two categories of characteristics: product categories and process categories. The product characteristics included are: constructs, models, methods, and implementations. The process characteristics included are: build, evaluate, theorize, and justify. This dissertation adapts their framework.

Table 4. Summary of Design Characteristics.

Characteristics	**Definition**
Constructs	vocabulary to describe concepts in the problem domain
Models	set of propositions expressing relationships between constructs
Methods	steps to perform a task
Instantiations	the outcome
Steps	
Build	creating an artifact to perform a specific task
Evaluate	the process of determining if progress has been made
Theorize	explain how the artifact and its interactions result in the observed performance
Justify	the generalization

The following are March and Smith's definitions for this framework which this research will adopt. 1. Constructs are a form of vocabulary to describe conceptualizations within a problem domain as the specifications of the solution space; they are a specialized language to that domain. 2. A model is a set of propositions or statements expressing relationships between constructs; in designing, models represent situations as problem and solution statements. March and Smith parsimoniously explain a model as a description of how things are. 3. A method is a set of steps to perform a task, such as an algorithm. 4. An instance is the realization of an artifact in its environment; instances demonstrate the feasibility and effectiveness of the models and methods they represent. To understand if an advanced analytic is effective, this dissertation is looking to the underlying notion of improving and expanding on existing methods that can be applied to healthcare data.

> Progress is achieved in design science when existing technologies are replaced by more effective ones. (March & Smith, 1995)

March and Smith proceed to define the process of research activities to include building and evaluating the artifact, and then theorizing and justifying the utility of the artifact. This research adopts the March and Smith definitions for these activities. 1. Building is defined as creating an artifact to perform a specific task, answering the question: does it work? 2. Evaluating is the process of determining if progress has been made, with the question how well does it work? Evaluation includes metrics that define what we are trying to accomplish. 3. Theorizing explicates the characteristics of the artifacts and its interaction with the environment into a synthesis of these characteristics, interaction, and environment. 4. Justifying is providing the explanation of this synthesis.

> Theorizing about instantiations may be viewed as a first step toward developing more general theories. Or as the specialization of an existing general theory. (March & Smith, 1995)

March and Smith define design science as the attempts to create things that serve human purposes, answering questions such as 'does it work?' and 'is it an improvement?'; Going from the 1995 paper to the 2013 paper, demonstrated is the evolution of these kinds of questions, as Gregor and Hevner have identified in their four classifications of problems. March and Smith go on to mention that design science, rather than posing theories, creates models, methods, and implementations that prove innovative and valuable, and that the process of this design is a theoretic contribution.

Table 5. March and Smith's Research Framework.

		Steps			
		Build	Evaluate	Theorize	Justify
Characteristics	Constructs				
	Model				
	Method				
	Instantiation				

This research adopts Hevner et al.'s design science guidelines (Hevner et al., 2004). These guidelines are referenced in table 6. These guidelines serve as the methodological basis to conduct design research, with each guideline resulting in a check on the practical and theoretical utility of the design in terms of validation.

Table 6. Hevner et al.'s Design Science Guidelines.

Hevner et al.'s Guideline	Dissertation Implementation
G1: Design science research requires the creation of an innovative, purposeful artifact (design as an artifact)	G1: Apply advanced analytics and machine learning algorithms to understand variances in cost and quality of care.
G2: Design science research requires a specified problem domain (problem relevance)	G2: Variance in the cost and quality of healthcare.
G3: The artifact must be evaluated	G3: Compare artifact results with best available standards, practice, and research.
G4: The artifact must either be novel, more efficient, or more effective (research contribution)	G4: Does this artifact provide a new way to understand variance in cost or quality? Is it faster? Does it explain more?
G5: The artifact must be rigorously defined (research rigor)	G5: Define how to apply algorithms to healthcare data in a manner that yields meaningful results.
G6: The process by which the artifact is created must be formed as a search process (design as a search process)	G6: Search for multiple algorithms and assess based on outcomes.
G7: The results must be communicated effectively (communication of research)	G7: Evaluate and communicate implications for healthcare practitioners and researchers?

Design Framework

Figure 2 provides the guiding research approach for this dissertation, synthesized from the review in this chapter. This approach enables the dissertation to take the three guiding research questions and generate sub-questions. This research approach is applied to these sub-questions as the process, with the product as the outcome of the approach. With the process and product, the result of each question is validated both internally and externally.

While Chapter 4 discusses the receiver operator characteristic curve as a mechanism for validation at length, the validation of this approach comes internally as a result that provides an alternative or expanded view of variance in healthcare data; external validation is applied by presenting the process as this approach for each sub-question, as well as the product from this approach, to industry experts to validate the process solves an existing problem.

This approach has the added benefit of incorporating the guidelines from Hevner et al.'s 2004 paper.

			Steps			
			Build	**Evaluate**	**Theorize**	**Justify**
Characteristics			G1	G3	G4	G5
Constructs	Kernel theory Ex: Heart-failure	G1				
Models	Meta-requirements Ex: k-means clustering	G2				
Methods	Meta-Design Ex: Quasi-experimental	G6				
Instances	Testable-Product Hypotheses Outcomes	G7				
			Design Method		Design-Process Hypotheses	

Figure 2. Design Framework.

Similarly, Walls et al.'s work provides this research approach with the notion of kernel theory, meta-requirements, meta-design, testable product-hypotheses, design method, and design process-hypotheses. These concepts form the basis for assessing the result of the design as a process and as a product, based on the product-hypotheses and they process hypotheses. March and Smith provide the terms constructs, models, methods, instances for the product, and build evaluate, theorize, and justify for the process. Built into each of these pieces of the construct is Hevner et al.'s guidelines as a check on the applicability of the research.

This approach can be applied in one of two ways. It can first examine the characteristics by following the steps of design science. An example of this first approach is that it can take the constructs associated with a question relating to heart failure, and follow through each step of the process until a realized instance is reached and justified. Similarly, it can take the characteristics through the whole process on a per-step basis. For example, it will fully identify the characteristics for design to build the design, and then follow through with evaluation, theorization, and justification. This dissertation utilizes the first approach; it will go through each step on a per characteristic basis. The benefit of assessing 'horizontally' through the framework is that each characteristic must pass both the design rigor as well as the theoretical rigor, and if it does not, then effort is not spent on a design that cannot be theorized.

The reviewed work in this section has been synthesized into an approach that meets all of the characteristics, components, steps, and guidelines for conducting research. This approach is applied the analytics reviewed in the next section.

Data Analytics in Information Systems

This dissertation is situated as exaptation design work, as illustrated in the previous section. To that end, it applies the existing solutions of advanced analytics and their underlying algorithms in new ways to healthcare data.

Data analytics has become a core business function (Kumar, Niu, & Ré, 2013). This function has evolved from efforts made in business intelligence, data mining, and analytics (Chen et al., 2012). Specifically, analytics in information systems are often researched by their function: explanatory, prescriptive, and predictive (Phillips-wren et al., 2015; Shmueli & Koppius, 2011). Given that their functionality is their key utility, this dissertation adopts these characteristics, as well as identifies additional characteristics of advanced analytics.

To understand and predict phenomena using healthcare data has found renewed motivation from incentives and penalties built into the ACA and enforced by CMS. These penalties apply to care providers for the readmission of patients within a 30 day window of certain chronic and expensive conditions. This penalty takes the form of a rate of 3% reimbursement penalty, an increase from 2% the year before, and 1% in 2013 (*Patient Protection and Affordable Care Act*, 2010). In the face of these penalties, healthcare providers are turning towards their own data to predict readmission risk. This motivates this dissertation to apply advanced analytics to healthcare data, and in doing so, assess the various established and latent factors around readmission in order to build alternative design models.

The application of analytics and data mining has been utilized by prior research efforts. One study built an explanatory model of healthcare quality using a decision tree and discovered important factors of inpatient mortality included stay, disease classification, discharge department, and age (Chae, Kim, Tark, Park, & Ho, 2003). Another study applied advanced analytics in the form of a multivariate regression to an EMR database and discovered a tie between knowledge management of the healthcare organization and the strategic guidelines in terms of cost (Abidi, 2001). In another study, the researchers applied classification and clustering analytics to understand what could be found in a data warehouse of diabetic medical information and discovered that the predictors of diabetes include age, sex, number of emergency department visits, office visits, comorbidities and heart failure codes (Breault, Goodall, & Fos, 2002). In another research study, pattern discovery was applied to a data set of 667,000 patients to understand a specific condition (Mullins et al., 2006). Lastly, studies have successfully utilized publicly available data and applied data mining in order to understand trends, such as Phillips-Wren et al's work mining lung cancer patient data in order to understand healthcare resource utilization; they did so by applying traditional regression on the public-use Medicare database of insurance claims (Phillips-Wren, Sharkey, & Dy, 2008).

While a tremendous amount of research effort investigating specific issues in healthcare data with specific techniques, this dissertation finds its novelty by providing a process of identifying the characteristics of the analytic as well as the characteristics of the data, and pairing the two together into a process and a product. The novelty manifests itself as a generalizable process that can be applied to numerous healthcare datasets for

any of the health conditions that are captured in the data, so long as the question conforms to the model requirements of the research approach.

These examples provide a basis to understand how analytics have been applied in the past. Of particular note is multivariate regression. Multivariate regression in the form of a generalized logit model is used by CMS and AHRQ on (YNHH-CORE, 2008) as the mechanism to assess heart failure readmission penalties. Because CMS and AHRQ use this model, healthcare provider entities refer to this as the standard model. The characteristics of multivariate regression are discussed in the next section, but this research is motivated to apply known solutions to new problems; multivariate is a known solution for a known problem. Given that this dissertation is motived to expand and improve the existing techniques for explaining variances in healthcare data, an alternative to the standard is presented. This search is conducted by examining the algorithms underlying the advanced analytics and comparing and contrasting the requirements and design.

The advanced analytics examined in this research have been established in prior research, and often find various tweaks and optimizations in the computer science literature. It is not the goal of this research to put forward a comprehensive state of the art knowledge classification of algorithms that power advanced analytics, but rather to build an understanding of the established algorithms and apply this understanding in terms of their requirements for instantiation to solve a class of open questions in the healthcare literature.

Algorithms of Advanced Analytics

Table 7. A Summary of Advanced Analytics' Characteristics.

	Explain	Predict	Prescribe	Visualize	Associate
Regression	X	X		X	
Decision Trees	X	X		X	X
Instance Based	X	X		X	X
Bayesian	X	X			X
Clustering	X			X	
Association Rule Learning	X		X		X
Artificial Neural Network	X	X			X

Advanced analytics are defined as multivariable and multidimensional statistical and mathematical models that generate statistical insight from data (Barton, 2012). The differentiating factor of advanced analytics from other classification of analytics lies within the complexity of the underlying algorithm (Lavalle et al., 2011). Sometimes referred to as data mining or machine learning, the underlying goal of advanced analytics is to take complex data and reduce the complexity (U. M. Fayyad, 1996; U. Fayyad, Piatetsky-Shapiro, & Smyth, 1996). While data mining sometimes includes specific techniques such as extract-transform-load (ETL), machine learning's core focuses on the specialized algorithms for transforming a cleaned and loaded data set into insight (Ayodele, 2010). This research uses the terms advanced analytics and machine learning interchangeably. Given that these algorithms are the underlying design of the advanced analytics, this section assesses the meta-design and meta-requirements of the advanced

analytics on a per-algorithm basis. These analytics can take the form of supervised learning, unsupervised learning, and semi-supervised learning (Ayodele, 2010; Carbonell, Michalski, & Mitchell, 1983; X. Zhu, 2007).

Supervised machine learning requires a training model for the data in order to generate insight from the data (Dasgupta, Sun, König, Bailey-Wilson, & Malley, 2011; Kotsiantis, Zaharakis, & Pintelas, 2007). This means that a small sample or example is provided to the algorithm and the user can specify the amount of accuracy of the result of the algorithm. The instances of problems that supervised machine learning addresses are classification and regression. Supervised machine learning includes multivariate regression, which requires a predefined model and confidence level.

Unsupervised machine learning takes an entirely unstructured input data set and builds models of similarity and dimensionality reduction (Niu, Zhang, Ré, & Shavlik, 2012). Typical instances of problems that unsupervised learning addresses include clustering and associations (X. Zhu, 2007). The implementation of unsupervised machine learning includes the Apriori algorithm, k-Means, and k-Nearest Neighbors clustering.

Semi-supervised machine learning takes a seed set of data and builds its own model from the seed data, using both structured data in the form of data being labeled as well as assigning its own meaning to the data (X. Zhu, 2007). Often used to structure data, instances of problems include classification of data. Semi-supervised often borrows from either supervised or unsupervised algorithms in order to achieve this goal of classification.

This research proceeds to identify several types of algorithms that power advanced analytics. It is not the goal of this research to provide a classification of the cutting edge of research as to what is new in machine learning and advanced analytics, but rather to build a foundation based on established algorithms. The underlying goal is understanding the output of these algorithms in a way that makes their model as it relates to design science applicable for constructs within the healthcare domain as it relates to building a better solution for exploring healthcare data. The outcome of this application will situate the constructs in the healthcare literature of cost and quality.

The algorithms that are reviewed include regression, decision tree, instance based, Bayesian, clustering, association rule learning, and artificial neural networks (Brownlee, 2015). Brownlee presents an overview of a variety of machine learning techniques, from which this research adopts the most applicable. While some of these algorithms fit into multiple categories, this research classifies them based on their yielded output, such that a clustering algorithm will identify latent clusters whereas a regression algorithm is meant to predict. This classification approach is adopted in an effort to avoid any confusion when considering the meta-requirements of the model for design.

Regression Algorithms

Regression algorithms in advanced analytics include ordinary least squares regression (OLSR), linear regression, logistic regression, and multivariate regression.

Prior research into advanced analytics and healthcare data typically utilizes regression as their primary source of inference. One example of this is modeling the

comorbidity and concentration of healthcare expenditure in heart failure patients as a regression model (Zhang, Rathouz, & Chin, 2003). Another example of regression utilizes logistic regression to understand the risk factors for complications and in-hospital mortality following hip fractures (Belmont et al., 2014); it should be noted that this research also employed a publically available data set of the National Trauma Data Bank. Additionally, examples of multivariate modeling can be found in the medical literature for identifying clusters for interventions (Rana, Gupta, Phung, & Venkatesh, 2015) as well as understanding healthcare provisioning on an economic basis (Lefèvre, Rondet, Parizot, & Chauvin, 2014).

This research uses regression to form the baseline expectation of results from the data, as published by AHRQ, since most healthcare literature adopts this analytical approach for their research. While the focus of this research is not entirely on regression, it examines several multivariate models in Chapter 3, based on knowledge presented from AHRQ. As this research seeks to address an instance of a problem in the form of predicting heart failure readmission and prescription utilization, it is examined based on the meta-requirements of what multivariate regression provides: statistical significance of correlations based on a provided model. This output informs the user of which data is related to which data, as well as if there exist any multicollinearity within variables, which can lead to more parsimonious models to yield the same insight.

Decision Tree Algorithms

Decision trees map observations to a targeted value in a predictive fashion. These algorithms include: classification and regression tree (CART), iterative dichotomiser 3 (ID3), and C4.5/C5.0.

CART is a decision tree algorithm that can examine categorical relationships or continuous outcomes (Speybroeck, 2012). CART has been used by Speybroeck with public health data in order to understand a binary health outcome of an individual based on whether that individual exhibits healthy or diseased factors. As the tree progresses, the homogeneity, or purity/accuracy of the tree increases at each node. Speybroeck demonstrates this by modeling education and income as nodes in the decision tree, and at each node, a decision is made if it meets a certain threshold. If the education requirement is not met, the then algorithm creates a branch that assesses the individual's income. If the income level is not met, the then algorithm creates a terminal node that the individual is going to be sick. Figure 3 shows Speybroeck's implementation of the CART algorithm using public health data.

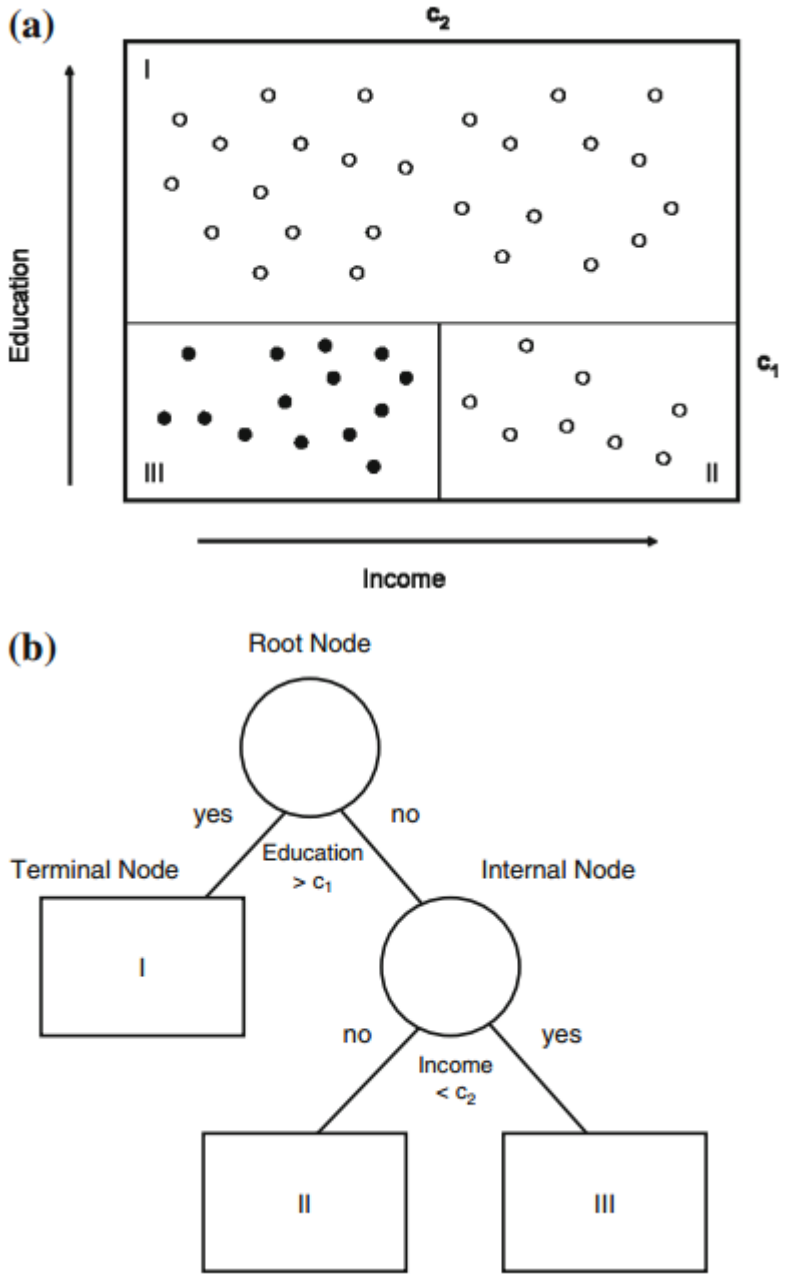


Figure 3. Illustrative Classification. (Speybroeck, 2012).

This provides a method for understanding not just how factors in the data are grouped together, but understanding how the dependencies of each factor influence each other. Since this algorithm requires some labeled data, it falls into a semi-structured approach.

Iterative dichotomiser 3 (ID3) is an algorithm used to generate a decision tree from a dataset (Quinlan, 1986). ID3 is the established algorithm and the precursor to the C4.5 and C5.0 algorithms (Singh et al., 2006). Typically used in natural language processing, ID3 creates branches of the tree by taking an original data set, assessing the attributes of the data, and iterating through the data based on entropy. With each iteration, the data is subset based on the attribute where the least amount of entropy is added to the system, or the most amount of information is added to the tree.

ID3 has been used in healthcare literature, examining decision trees for monitoring cardiac complications of diabetes (Kelarev, Abawajy, Stranieri, & Jelinek, 2013). This particular research tested various different decision tree classifiers against each other and found that while CART provided the best performance against a historical data set, ID3 was second in performance and error rates. ID3 provides a validated algorithm for applying decision tree machine learning to the healthcare claims data set. In doing so, it requires input of data as well as attributes of the data, which makes it a semi-structured approach. The outcome of the algorithm will provide a tree structure of how the data relates to each other based on their attributes.

Quinlan extended ID3 into the C4.5 algorithm to include normalization of information gain at each decision split (Quinlan, 1993). Additionally, C4.5 can handle

data with missing attribute values or attributes with different weights. C5.0 is the commercial implementation of C4.5.

Decision trees have been examined for knowledge discovery in the information systems literature by examining the tools for analysis and visualization of web mining (Chung, Chen, & Jr, 2005). Using this identified value, this dissertation adopts those attributes and characterizes its ability in terms of healthcare data as it can explain, prediction, as well as associate. On top of this, the visual representations of the trees and clusters make it valuable and versatile for design. By using total cost of a claim as the final terminal node, it learns from the data depending on which factors and creates branches of the tree that provide an alternative view of healthcare data that improves and expands the understanding of variances based on these branches.

Instance Based Algorithms

Instance based algorithms take a model decision problem or exemplar and create a database of associations (Daelemans and Bosch, 2005). These models are sometimes referred to as memory-based learning. Instance based algorithms include k-Nearest Neighbor (kNN) and Self-Organizing Map (SOM).

kNN is an algorithm used for both classification as well as regression (Altman, 1992). The kNN algorithm can output class membership a well as the value of an object as it relates to its nearest neighbors. Altman specifies the kNN algorithm as a lazy algorithm, where it approximates values to a local set and defers computation until classification. The algorithm itself first trains on a sample vector and class label, and then

produces a Euclidean distance of how far apart each classification is. The kNN algorithm struggles with higher dimensionality models, and Altman cautions that dimensionality reduction may need to occur before employing the algorithm. This creates a design opportunity to use the results of one algorithm, such as principle component analysis, as the inputs for a new instance based algorithm. By leveraging the characteristics of a regression algorithm, the algorithm can improve the benefits of the instance based algorithm.

For this research's datasets, the kNN algorithm produces the distance of each classifier in the dataset, which can then visually represent their adjacency or distance. To exemplify how this research will employ this algorithm, consider again the question of understanding heart failure readmission and the factors that contribute to readmission; kNN creates a model that shows in an Euclidean plane the distances and adjacencies of each healthcare metric.

Similar to kNN, the self-organizing map algorithm creates a low-dimension representation of data in the form of a map (Kohonen, 1982). It requires a training dataset as well, in order to build its model, and then maps the model while automatically creating new class vectors. Each node carries a weight as well as a vector. While SOM is considered an artificial neural network, it operates on a per-instance basis. The goal of SOM is to create a network that has the same response as the training set; by taking a small sample in the training set, it attempts to use that model on the large data set and assesses the accuracy of both the training and output, and assesses the 'fit' of the output against the training. The value of this particular algorithm is improving our current

understanding of factors as they relate to cost and quality of care by assessing current models against a larger data set. SOM has been demonstrated in Journal of Management Information Systems as well to be an effective advanced analytic for data mining risk groupings of prostate cancer patients (Churilov, Bagirov, Schwartz, Smith, & Dally, 2005).

The value that can be gained from instance based algorithms comes from its ability to map distances of factors in Euclidean space. If we want to better understand associations of data, this algorithm provides an empirical assessment of the distances of these factors in the form of a map.

Bayesian Algorithms

Bayesian algorithms explicitly apply Bayes theorem (Bayes, 1763) as a probabilistic event. This research examines the applicability of Naïve Bayes as well as Bayesian belief networks.

The Naïve Bayes classifier is an algorithm that functions on Bayes theorem by applying a strong independence of probability assumption between the items being classified (Rish, 2001). This algorithm has been applied in the past in medical diagnosis of conditions, assessing the probability of a condition given independent assumptions of other health conditions (Lakoumentas et al., 2012; Matwin & Sazonova, 2012). The Naïve Bayes algorithm is of interest when considering the readmission and pharmaceutical problem; naïve Bayes provides a way to assess readmission prediction independent of other assumptions as well choice of pharmaceutics. It provides a

mechanism by which this research can model a readmission true-positive, or a generic drug true-positive.

A Bayesian belief network functions in a similar manner, except that it functions in an semi-supervised manner and the result is the probabilistic relationships between the data. A Bayesian belief network can also be used for variable elimination and dimensionality reduction (Cheng, Bell, & Liu, 1997). Building on this, Bayesian belief networks as a function of a computer-assisted diagnosis of breast cancer, where the researchers employed the algorithm to greater accuracy than logistic regression by itself (Wang, Zheng, Good, King, & Chang, 1999).

A Bayesian approach to machine learning provides this research with a set of tools to create probabilistic models. This research benefits from using probability when considering the factors that comprise readmission, cost, and pharmacy decisions from the point of view of the patient. Given that these models are run on historical data, the results of the beliefs of the theorem are then validated against what actually occurred. This provides the research a mechanism for testing the applicability of Bayesian inference on claims data.

Clustering Algorithms

One of the most established mechanisms of machine learning and advanced analytics is the class of algorithms known as clustering algorithms (Guzella & Caminhas, 2009; A. Jain, Murty, & Flynn, 1999). These clustering techniques have found utility in a variety of spaces, from optimizing locations data in data warehouses (el Mensouri, Beqali

El, & Elhoussaine, 2013) to identifying clusters of factors that contribute to healthcare utilization and conditions (Lefèvre et al., 2014; Rodolfo, Pérez-ortega, Miranda-henriques, & Reyes-salgado, 2010). This research examines k-Means, expectation maximization, and hierarchical clustering.

k-Means clustering is a mechanism for partitioning observations into *k* number of clusters (A. K. Jain, 2010; MacQueen, 1967). k-Means identifies the means of data iteratively and identifies spatial local optimums. As it iterates, the algorithm refines the location of each means as it creates associations of each observation to the nearest mean. Centroids are assessed and refined until the algorithm reaches convergence. It has been demonstrated that the result of k-means clusters can be visualized using Voronoi diagrams (Burkey, Bhadury, & Eiselt, 2011). This visualization can take the result of k-means and either optimize to local centers, or optimize for maximum distances. This is of particular relevance for this research, as it seeks to examine outliers in a massive data set, and examine the effect those outliers have on modeling the data. While there are variations and extensions to the k-means algorithm, such as k-medians, and k-means++, this research employs the established k-means algorithm.

Expectation maximization (EM) is another iterative clustering algorithm that identifies maximum likelihood estimates of parameters by using a model that depends on unobserved latent variables (Dempster, Laird, & Rubin, 1977). EM functions by taking an observation data set and generating an observed latent set to approximate the likelihood of an observation. The utility of EM for this research comes from its ability to predict based on an observed data set and identify latent variables from those

observations; this relevance is rooted in the readmission problem. EM provides an algorithm for identify factors that may yet be unknown in the readmission of heart failure calculus.

Hierarchical clustering (HC) is a clustering algorithm that builds hierarchies of clusters (Johnson, 1967) (Rokach and Oded, 2005). Hierarchical clustering can take the form of agglomerative, which is a bottom up approach in that it starts with each observation as its own cluster and merges the clusters through iterations, or it can be divisive, which treats the data as one cluster and partitions through each iteration. The results of HC often take the form of a dendogram visualization (Langfelder & Horvath, 2012).

While this research has identified three separate methods for clustering the healthcare data, the meta-design is very similar while the requirements are different; the design is to cluster data, while the way in which the clusters happen differ. The differentiation between k-means, EM, and HC for this research manifests itself in two ways: 1. The meta-requirements of each algorithm in terms of their mechanisms for identifying clusters of factors in healthcare data, which builds to 2. the ways in which these clusters can be visualized as an instance of the design. It has been shown that the problem is not the implementation of advanced analytics, but rather the interpretation (Davenport, 2013) and visualizations provide a powerful mechanism for the end user to interpret the result of the algorithms (Andrienko & Andrienko, 2013; B. Zhu & Watts, 2010).

Association Rule Learning Algorithms

Association rule learning algorithms are algorithms that build association rules that best explain observed relationships between variables in the data (Piatetsky-Shapiro, 1991). The result of Piatetsky-Shapiro's research is the identification of association of strong rules in large-scale transactional data. Agrawal et al introduced the concept of strong rules via example that if a customer buys both onions and potatoes, there is an association that they will buy hamburger meat (Agrawal, Imieliński, & Swami, 1993). Association rule learning algorithms have been established in two forms: apriori and éclat.

The apriori algorithm is a set mining algorithm meant to be employed on transactional databases (Agrawal & Srikant, 1994); it should be noted that the proprietary data that this research utilizes comes from a transactional database. It functions by identifying the frequencies of individual items in a database and extends them into larger item sets, as long as the item sets occur often enough in the data. Apriori is defined as a breadth-first search algorithm as it counts the occurrences of support in itemsets and employs a candidate generation function. This algorithm has been used to identify trends in a database (Agrawal & Shim, 1996; M. J. M. Zaki, Parthasarathy, Ogihara, Li, & Others, 1997). The benefit of this algorithm for this research is twofold: the data exists as a transactional database, so no ETL processes need to be applied, and the result of the apriori algorithm can be replicated across any publically accessible claims database to test and validate the results.

If the apriori algorithm is a breadth-first search function, the eclat algorithm is considered a depth-first search algorithm (M. J. M. Zaki et al., 1997; M. J. Zaki, 2000). Eclat is an acronym for equivalence class transformation. Whereas apriori looks at itemset count, eclat looks at itemset intersection.

By utilizing association rule learning, this dissertation gains aspects of design for understanding connections in the data that may not have been established in the vast array of medical or healthcare literature. Moreover, this dissertation provides a method and results of association rule learning to healthcare data.

Artificial Neural Networks

Referred to as neural networks in the information systems literature, it is sometimes referred to as artificial neural networks in other fields, these classes of algorithms mimic the function of a biological neural network and back propagation (Rumelhart, Hinton, & Williams, 1986). While there is a vast array of techniques being applied in machine learning, one that stands to benefit this research is that of back-propagation artificial neural networks. Back propagation has two steps: forward propagation of the training through a neural network in an effort to create outputs, and then backwards propagation of the model using the neural network to create deltas (LeCun, Bengio, & Hinton, 2015). Since backwards propagation assesses the error rate of the neural network, this research employs the backwards propagation to the healthcare data in order to understand the associations and errors of those associations.

Summary of the Design and Requirements for Analytics

This chapter has presented a review of the components of design science research as well as broken down advanced analytics into their underlying algorithms, their design and their requirements. Informed by the notion that design science contains both a product as well as a process, and that these two must contain a meta-design and meta-requirements, this chapter has assessed the viability of established algorithms that power advanced analytics, and presented an approach to apply design science and machine learning to data. The next chapter presents the models, methods, and constructs from the problem domain.

CHAPTER III

A REVIEW OF CONSTRUCTS, MODELS, AND METHODS

This chapter reviews the work that has been conducted in healthcare as it relates to cost and quality of healthcare for heart failure. The purpose of this chapter is to review existing knowledge of cost and quality. This review informs the research question as it relates to improving and expanding understanding variance in cost and quality of care. Specifically, this chapter outlines the constructs, models, and methods for building an instance of a problem to answer the research question.

The structure of this chapter is as follows. First, cost is examined in the context of its constructs, models, and methods. Following the same structure, quality measures are examined. From these general forms of understanding, this research then examines open research issues as they related to heart failure readmission as a quality construct. From these reviews, this dissertation identifies the opportunity to improve design in the form of creating an instance of modeling heart failure readmission. Following this, the data being used to build this instance is described.

Examining Cost

When examining the cost of healthcare, literature (Hey, 2010; Takeda et al., 2012; Zhang et al., 2003) has examined the relationships of diagnosis, cost of treatment, economic burden of the individual, and health conditions of the individual. To make this

research viable, the studies focus on one specific health condition and posit models around contributory factors with cost as the dependent variable. These factors can include diagnosis, condition as chronic verses acute, inpatient vs outpatient, medication usage, comorbidities, weekday verses weekend of admission, point of admission, and demographic controls.

For example, one study assessing the economic burden of COPD to the USA examined demographics, medications, level of education, and healthcare costs per visit and per treatment (Britton, 2003). Another study assessed the economic burden of cost of a surgical acquired infection and found contributory factors to the cost to include assessing the cost of diagnoses after the patient is discharged as well as the opportunity costs of lost bed-days for the hospital (Graves et al., 2008). Another study examined the cost of healthcare as it relates to medication adherence on hospitalization risk and found associations with diabetes and hypercholesterolemia medication adherence lowered the overall cost of care (Sokol, McGuigan, Verbrugge, & Epstein, 2005). In this stream of literature, there are also review efforts of contributory factors to the cost of healthcare, such as the cost of direct medical costs for overweight and obese individuals in the United States (Tsai, Williamson, & Glick, 2011); results of studies like these include recommendations as to which factors should be included in the representative samples as well as the prediction models; for example Tsai et al., 2011found that BMI cutoff information contributes to improving the prediction models when assessing the cost of care.

Ultimately, cost is an easy construct to understand abstractly, however defining the cost of care is a harder problem. Identifying the factors that comprise diagnoses and treatments is an ongoing and constantly shifting effort. This research contributes to that effort in providing the results of applying advanced analytics onto healthcare claims in order to identify and expand on our understanding of cost and quality. For this effort, this dissertation takes the two datasets, as well as publically available census data, and combines the data sets to build models that go beyond the traditional multivariate prediction models that incorporate the known factors that contribute to cost. For this dissertation, socioeconomic and geographic dispersion is accounted for when considering the problem of identifying an individual at risk of a heart failure readmission and how it relates to cost of care provided; this is a direct answer to the call for this effort in the Cochrane Collaboration, a publication that assesses over 2,000 factors that affect the clinical service provided for heart failure patients (Takeda et al., 2012) and concluded that multivariate models could benefit from socioeconomic factors.

Examining Quality of Healthcare

While cost as a construct is somewhat easy to understand, defining quality of healthcare adds a layer of complexity due to its abstract nature. This effort, in the United States, is guided by AHRQ with their four quality indicator categories: Prevention Quality indicators, Inpatient Quality Indicators, Patient Safety Indicators, and Pediatric Quality Indicators. Each of these sets of quality indicators are used as the industry

benchmark against what constitutes quality in healthcare (Serra-Sutton, Serrano, & Carreras, 2013; Stelfox, Straus, Nathens, & Bobranska-Artiuch, 2011).

AHRQ - Prevention Quality Indicators

Prevention quality indicators (PQIs) are a set of measures that can be used with hospital inpatient discharge data to identify quality of care for "ambulatory care sensitive conditions," (AHRQ, 2015e). These indicators are used for assessing outpatient care to prevent a readmission or hospitalization when an intervention can be identified to prevent readmission. PQIs are used as a screening tool in order to identify health care problem areas. These indicators include assessing care issues such as diabetes short-term complications admission rate, COPD in older adults' admission rate, low birth weight rates, as well as heart failure admission rates. This research further examines the PQI for heart failure admission rates.

AHRQ provides the PQI #08 indicator to represent the heart failure admission rate (AHRQ, 2013b). The outcome of this quality indicator takes the shape of a rate of admission. From AHRQ, this PQI is described as 'Admissions with a principle diagnosis of heart failure per 100,000 population, ages 18 years and older. Excludes cardiac procedure admissions, obstetric admissions, and transfers from other institutions.' The PQI is calculated by taking the discharges from a hospital with an ICD-9 code for heart failure and dividing it by the population of 18 years and older in a metropolitan or county area. This provides this dissertation with a baseline for guidance as to which ICD9 codes to use when employing the advanced analytics on the healthcare claims dataset, as well as

an understanding of the current benchmarks being employed as to how quality is considered a construct in healthcare. The outcome if this PQI is a ratio, and while the ratio can represent the number of discharges to total population, it is deficient in providing information on an individual level.

AHRQ - Inpatient Quality Indicators

Inpatient Quality Indicators (IQIs) are the set of measures that provide a perspective of hospital quality of care using hospital administrative data (AHRQ, 2015b). These indicators are meant to reflect the quality of care provided inside hospitals and include inpatient mortality for defined procedures and medical conditions. Specifically, it seeks to examine underuse, overuse, or misuse of hospital procedures depending on mortality rates. These quality measures identify problem areas in hospitals that need further examination, based on discharge records, mortality rates, and hospital transfers. Examples of IQIs include coronary artery bypass graft volume, acute myocardial infarction mortality rate, heart failure mortality rate, and hysterectomy rate.

AHRQ provides IQI #16 indicator to represent the heart failure mortality rate. The outcome of this quality indicator is a rate based on the in-hospital deaths per 1,000 discharges with heart failure as a principle diagnosis for patients ages 18 years and older (AHRQ, 2013a). This rate is calculated by taking the number of deaths that meet the inclusion and exclusion rules as the numerator, and dividing by discharges for patients ages 18 and over with an ICD-9 code corresponding with heart failure.

AHRQ - Patient Safety Indicators

Patient Safety Indicators (PSIs) are a set of indicators providing information on potential in hospital complications and adverse events following surgeries, procedures, and childbirth (AHRQ, 2015c). These indicators were developed in order to help hospitals understand and identify potential adverse effects as well as assess incidents that may affect patient safety. Examples of PSIs include retained surgical item count, postoperative sepsis rate, and death rate among surgical inpatients with serious treatable conditions.

AHRQ provides PSI #04 indicator to represent the death rate among surgical inpatients with serious treatable complications (AHRQ, 2015a). The outcome of this indicator is a score that represents the in-hospital deaths per 1,000 surgical discharges, among patients 18-89 with serious treatable complications (sepsis, shock, acute ulcer). To calculate this indicator, the numerator is defined as the number of deaths among cases meeting the inclusion or exclusion use rules for the denominator. The denominator is defined as surgical discharges, for patients 18-89, that has any ICD-9 procedure code and has a principle procedure occurring within 2 days of admission and meets the inclusion and exclusion criteria. Table 8 provides an overview of the criteria for this quality indicator.

Table 8. AHRQ PSI #04 Stratum Criteria.

Stratum	Condition
Stratum 04A	Deep Vein Thrombosis/Pulmonary Embolism
Stratum 04B	Pneumonia
Stratum 04C	Sepsis
Stratum 04D	Shock/Cardiac Arrest
Stratum 04E	Gastrointestinal Hemorrhage/Acute Ulcer

AHRQ - Pediatric Quality Indicators

Pediatric Quality Indicators (PDIs) function as the fourth set of quality indicators put forth by AHRQ. This set of indicators is meant to assess problems that pediatric patients might experience being exposed to a healthcare system and identify prevention mechanisms (AHRQ, 2015d). This research acknowledges this set of quality indicators as being put forth by AHRQ but is not the primary focus of this research.

Constructs, Models, and Methods of Heart Failure

Heart failure readmission rates have become a quality measure for hospitals. In this section, this dissertation examines systematic reviews of literature that have examined the constructs and models that make up heart failure, as well as puts forward several open questions from these reviews. While these reviews focused on heart failure readmission, they also identify opportunities for organizations to benefit from understanding the factors that comprise cost and quality. Heart failure readmission is examined because it is a quality measure that also has direct cost implications.

To begin, a 2013 review of social factors on risk of readmission in mortality assessed 72 papers examining mortality of individuals with community-acquired pneumonia (CAP) and heart failure (Calvillo–King et al., 2013). This review was conducted by building a model that included: age, gender, race, education, employment, income, socioeconomic status, type of insurance, risky behaviors such as smoking, alcohol, and dieting. Calvillo-King et al.'s review of these 72 papers utilization of analytic models found that 52 of the papers used multivariate regression, 12 used Cox proportional regression, the remaining used univariate or bootstrapping as their modeling tool. This review found that older age is the most consistent risk factor. The study also found that low income, education, and Medicaid were predictors of risk, and that neighborhood such as rurality and distance to hospital were also predictors of post-hospital outcomes. Calvillo-King et al., 2013 concludes the review by making the call for future research in the areas of social factors on readmission and mortality by leveraging the growing amount of digital data that is available. This dissertation is informed from this effort in that it is demonstrated efficacious to pair social dimensions with heart failure to examine both readmission risk as well as mortality. From a cost and quality perspective, this has direct implications on the inputs to the quality indicators from AHRQ, as well as determining patients at risk of readmission.

In another 2011 review, the authors examined validated readmission risk prediction models, assessed the models performance, and assessed 30 studies with 26 unique models as well as the models' applicability for clinical and administrative use (Kansagara et al., 2011). This review assessed One of the findings of Kansagara et al.'s

review is that only one of the models in their review attempted to explicitly define and identify potentially preventable readmissions. Fourteen of the remaining models were retrospective, and seven were used for identification of high risk patients. The remaining five models looked at factors at the point of discharge. The model that predicted potentially avoidable readmissions found that complications of surgical care, complications of nonsurgical care, drug-related adverse events, missing or erroneous diagnosis, and premature discharge were the leading factors (Halfon et al., 2006). The ultimate outcome of Kansagara et al.'s review was that most readmission risk prediction models perform poorly and that more factors need to be assessed. The factors that Kansagara et al., 2013 identified in their review include: medical diagnoses, mental health comorbidities, illness severity, prior use of medical services, overall health, sociodemographic factors, and social determinants of health including income, insurance status, education, marital status, and access to care. This review proposed that hospital and health system level factors, such as timeliness of post-discharge follow-up and quality of medication reconciliation could contribute to heart failure readmission risk.

A 2014 review of risk prediction in patients with heart failure examined 64 models (Rahimi et al., 2014). This specific review focused on models that predicted death, hospitalization, or hospitalization and death. This review found that the strong predictors of death included: age, renal function, blood pressure, blood sodium level, left ventricular ejection fraction, sex, brain natriuretic peptide level, diabetes, the New York Heart Association functional class, body mass index, and exercise capacity. All 60 models were multivariate linear or multivariate hierarchical linear models. This review

did find that models that focus on both hospitalization and/or death have a lower discriminate ability than models that focus on just death or just hospitalization. This review proposed that the tools being used to predict hospitalization or death from a heart failure event are being underutilized or mis-utilized and that further research should be put into making these models more accessibility to clinicians and researchers.

A 2015 paper conducted a systematic review of the association between quality of hospital care and readmission rates for patients with heart failure by identifying the American College of Cardiology and American Heart Association quality measures (Fischer, Steyerberg, Fonarow, Ganiats, & Lingsma, 2015). These quality measures from the ACC and AHA include angiotensin-converting enzyme inhibitor/angiotensin receptive blocker use, evaluation of left ventricular systolic function, smoking cessation counseling, discharge/compliance instructions, and anti-coagulant at discharge usage. This review examined 18 models, and found limited support for each quality measure. This study suggests that cost and readmission may be influenced by patient characteristics such as patients' life circumstances and nature of posthospital. This review concluded summarizing that readmission rates may not be affected by evidence-based ACC/AHA in-hospital process indicators; it calls for more research to be conducted examining if in-hospital quality of care is a determinant of readmission, or if readmissions are influenced by post discharge care.

From these four systematic reviews, this dissertation has identified existing constructs, models, and methods being used in the study of heart failure readmission. Additionally, opportunities exist from each of these reviews. Calvillo-King et al. several

models that include socioeconomic factors, but makes a call to leverage more data in more ways (Calvillo–King et al., 2013). Kansagara et al. suggests that examining the time between discharge and follow-up would prove insightful (Kansagara et al., 2011). Rahimi et al. identify that heart failure readmission risk could benefit from a variety of models that could be employed, as the current models are being underutilized (Rahimi et al., 2014). Fischer et al. suggest that quality measures may be more indictivative of a hospital's performance rather than that of an indivdual, and that the relationship between a patient's life circumstance should be examined in the light of readmission (Fischer et al., 2015). This review of literature builds the basis for examining cost and quality of care within the class of problems of heart failure. Situated within this class of problems are the instances of AHRQ quality constructs, as well as that of heart failure readmissions which spans both cost as well as quality. The next section of this dissertation build upon these instances.

Forming Instances to Address the Gaps

The research question driving this dissertation is: how can we improve and expand upon our current understanding of the variances in the cost and quality of care for specific health conditions. This question has been scoped to examine a specific health condition of heart failure. This dissertation has also examined cost and quality of healthcare, as well as the related constructs, models, and methods that comprise cost and quality when considering heart failure.

Specifically, work has been conducted around the ideas of quality measures and adopted by AHRQ, as well as the notion that readmission of a patient with heart failure is considered a quality measure for a hospital and a quality measure for an individual. This positions the research framework to examine the chronic conditions of heart failure by examining the costs and quality metrics associated with heart failure. Two instances that will be examined include heart failure readmission, as well as AHRQ metrics. Given the cost incentives, this research further examines heart failure readmissions in 30, 60, and 90 day windows.

Inherent to the design of this dissertation is an answer to the gap that Rahimi et al., 2014 identify for future research, by the design of using advanced analytics to address the problem of heart failure readmission. While the reviewed literature examines multivariate models, and the access to these models is, as suggested, not being used, this dissertation employs advanced analytics as the modeling technique. Each of the advanced analytics from Chapter 2 are operationalized into instances to explain and predict heart failure readmissions in new ways. The work by Calvillo-King et al., 2013 puts forward the constructs required in order to assess heart failure based on social aspects, including sociodemographic and socioenomic aspects. More recent work in the field of information systems has examined the applicability of predictive analytics for readmission of patients with congestive heart failure and demonstrated that their modeling techniques generated more accurate results than the normally employed logit regressions (I. . Bardhan, Oh, Zheng, & Kirksey, 2015). Additionally, work has been conducted using supervised machine learning to predict length of stay to prioritize discharges within the realm of

healthcare (Barnes, Hamrock, Toerper, Siddiqui, & Levin, 2015). Similar to this work, this dissertation assesses multiple types of advanced analytics against instances of the problem of cost and quality of care for patients with heart failure.

The instances of a problem this dissertation is addressing is the following: assessing the most significant factors in the calculation of quality measures, assessing variances in cost of treatment for patients with heart failure and identification of the most significant factors, and identification of an individual at risk of a heart failure readmission to a hospital. These problems can result in death for the individual, results in fines for the hospital, and results in increased costs for both. To address this problem, this dissertation applies machine learning as advanced analytics to build models for this identification.

CMS Logit Model

The CMS logit model the standard prediction model used by CMS and AHRQ to predict heart failure readmissions. Because CMS and AHRQ use this model, healthcare provider entities pay attention to this model and use it as their baseline for predicting patients at risk of a heart failure readmission.

The logit model is a hierarchical model that predicts readmission as a function of a patient's diagnoses, age, sex, and hospital at which they were treated. The diagnoses are grouped into an hierarchy using condition category codes. Condition category codes are further explained in Chapter IV. Condition category goes have the added benefit of automatically adjusting for risk factors.

The CMS logit model attempts to create a construct known as readmission. The model uses the factors of demographics and condition categories to model this construct. The model generates a hospital-specific prediction, which accounts for the predicted quality of healthcare as measured by heart failure readmission.

The generalized logit model follows: $h(Y_{ij}) = \alpha + \beta Z_{ij}$ where *i* denotes the hospital and *j* denotes the patient. Z is defined as the specific set of patient covariates, modeled by condition category codes. The logit function is specified as follows: $logit\left(P(Y_{ij} = 1)\right) = \alpha_i + \beta Z_{ij}$ where $\alpha_i = \mu + \omega_i;\ \omega_i \sim N(0,\ t^2)$ denotes the logit function if the patient was admitted. This specific function is modeled replicated in Chapter IV.

Summary of Chapter

This review of the literature has presented and contextualized cost and quality in the domain of healthcare. To better understand the work that has been put forward in cost and quality of care, this chapter examines the specific condition of heart failure, and how it relates to quality of healthcare; this chapter assembles the constructs, models, and methods as they exist in the current healthcare literature. From this examination, a specific instance to answer the research question is generated. Chapter IV provides a discussion of the data and a replication of the CMS logit model.

CHAPTER IV

DATA ETL AND CMS LOGIT REPLICATION

This chapter describes the data being used for this dissertation as well as the extract-transfer-load (ETL) process for the dataset. Additionally, this chapter examines an established thirty-day all cause readmission model for patients suffering from heart failure. The established model which is examined in this chapter is the model that the CMS has adopted for readmission rate predictions; given that CMS is the entity that sets both quality measures as well as penalizes hospitals, this model is valuable in order to establish a baseline prediction which will serve as a comparison for advance analytics. The purpose of this examination is to take an established model for predicting readmissions and establish a baseline by which alternatives can be explored. This comparison is achieved by calculating a confusion matrix of actual against predicted outcomes, which is also explained in this chapter. This chapter replicates this established model and provides the baseline in the form of summary statistics around the receiver operator characteristic (ROC) curve.

This chapter defines the fields in the data and the process by which the data was extracted, transformed, and loaded for analysis in this dissertation. Finally, this chapter replicates the established readmission model and provides the results of this model as a baseline for comparison. The structure of this chapter is as follows. First, all relevant fields in the North Carolina State Health Plan (NCSHP) are discussed. Next, the ETL

process by which the master database is filter is presented. This chapter then replicates the CMS readmission model.

NCSHP Data

The NCSHP data is a proprietary data set which accounts for all healthcare claims for individuals that subscribe to the healthcare plan offered by the state of North Carolina. This data set spans June 2007 to December 2014 and is contained in four source SAS flat files separated by claim process year. These files contain every insurance claim for every member subscribed to the plan in this time frame. The files are separated into two types: member eligibility and member claims. The number of eligible and active members subscribed to the plan is approximately 700,000 individuals while there are 931,302 unique member IDs in the dataset. This member ID discrepancy is the result of employee turnover, marriages that combine health plans, and deaths. The number of claims total 330,967,534 claim lines, where any insurance claim can have multiple claim lines; every record represents a line of a claim, with the average claim having 4.3 claim lines. The eligibility records total 6,608,077 records; these 6.6 million records provide the history of eligibility for the 931,301 member IDs at any given time, as well as the plan to which they were subscribed.

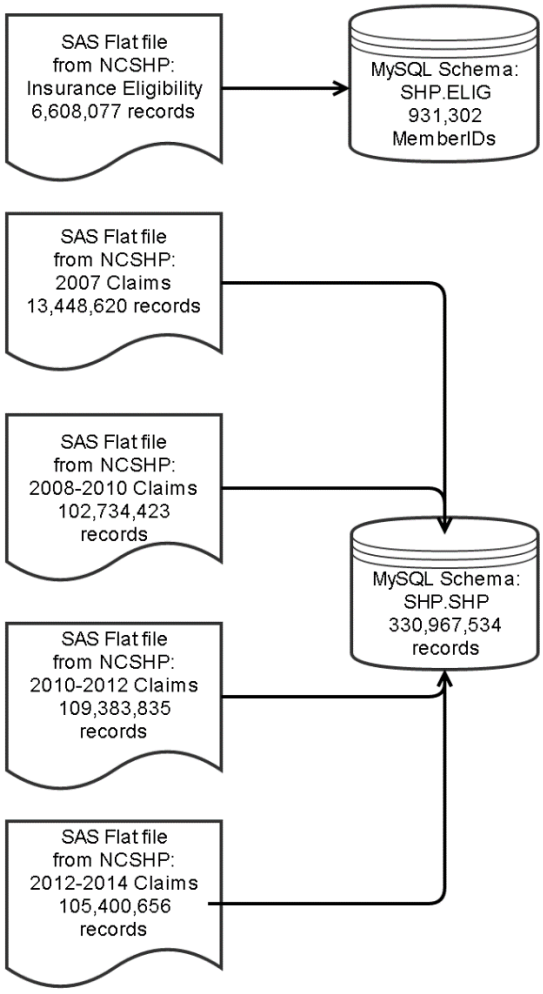


Figure 4. Data Sources.

The NCSHP dataset is classified as a limited data set by the Health Insurance Portability and Accountability Act of 1996 (HIPAA). HIPAA defines a limited dataset as: *a limited data set excludes specified direct identifiers of the individual or of relatives, employers, or household members of the individual* (Public Law 104-191, 1996). While individually identifying information is not available in this dataset, it is possible to identify an individual as the data represents claims for a person. The fields include ICD-9 diagnosis codes, the codes that medical practitioners use to numerically represent a diagnoses and comorbidities of patients. While there are over 100 fields in the dataset, analysis is confined to the fields found in table 9.

Table 9. NCSHP Data Fields.

NCSHP Fields	**Definition**
Member ID	Unique identifier of a patient
Date of Birth	Date of birth
Member gender	Biological sex of the patient
Service Start Date	Day the claim service started
Service End Date	Day the claim service ended
Diagnosis 1	ICD-9 code used to denote the primary diagnosis being treated
Diagnosis 2	Secondary diagnosis of the claim
Diagnosis 3	Third diagnosis of the claim
Diagnosis 4	Fourth Diagnosis of the claim/comorbidity
Diagnosis 5	Fifth Diagnosis of the claim/comorbidity

Extraction of Data

Given that the purpose of this dissertation is to study techniques that are relevant to examine heart failure readmissions, the first step is to identify in the data the patients who have experienced a heart failure or had a coded history of heart failure in any of the diagnosis columns. The next step is to identify those patients' history of hospital stays known as their inpatient history.

To achieve this, the member ID was used to pull the entire history for each patient who had a coded heart failure. The resulting dataset provides a complete claims history of every patient with heart failure, including a history of their diagnoses before their initial admission, for their admissions and readmissions, and after their readmissions. The procedure followed to identify this pool of patients follows.

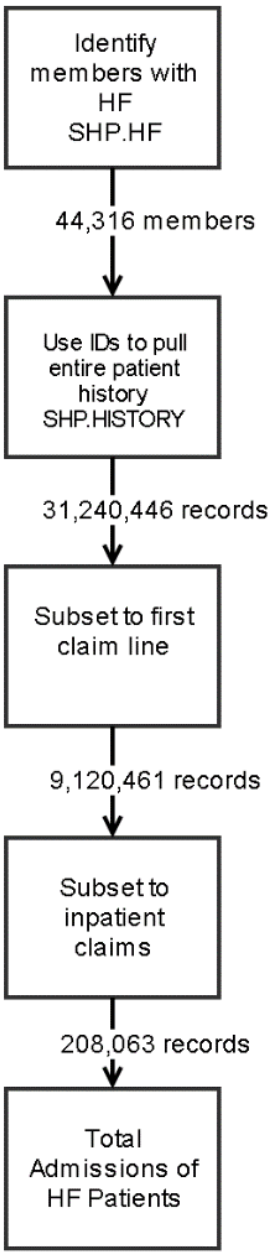


Figure 5. Identification and Extraction of HF Members.

The source data files containing the 330 million records were loaded into a MySQL database. By loading the data into a database, this allows for replication of code and analysis, as well as transparency in each step.

Step 1 – Identifying HF Admits/Readmits

The first step was to identify claims of heart failure patients. To do so, the inclusion criteria of Preventative Quality Indicator #08 from AHRQ was used. Any record that had one of the corresponding ICD-9 codes, as referenced in table 10, for any of the diagnosis columns in the source data, was loaded into a new table. Using the member IDs of patients with a heart failure diagnosis, the master database was queried to retrieve all claims for those. The resulting table provides the entire claim history for the heart failure patient, including before and after a heart failure inpatient claim. The total number of records in this table is 31,240,446 with 44,316 unique patients. The resulting table provides a baseline for generating a member's history for 90 months of data.

Table 10. ICD-9 Inclusion Criteria.

ICD-9-CM	Description
398.91	Rheumatic Heart Failure (congestive)
402.01	Malignant hypertensive heart disease with congestive heart failure (CHF)
402.11	Benign hypertensive heart disease with CHF
402.91	Hypertensive heart disease with CHF
404.01	Malignant hypertensive heart and renal disease with CHF
404.03	Malignant hypertensive heart and renal disease with CHF and renal failure (RF)
404.11	Benign hypertensive heart and renal disease with CHF
404.13	Benign hypertensive heart and renal disease with CHF and RF
404.91	Unspecified hypertensive heart and renal disease with CHF
404.93	Hypertension and non-specified heart and renal disease with CHF and RF
428.xx	Heart failure codes

Step 1a – Filter Duplicate Claims via Claim Lines

An encounter can contain many claims, and each claim will have its own information. The claim lines provide additional information around billing and the adjudication progress. When filtering the adjudication columns from the data, this created duplicate columns of diagnoses and service start/end events. The history table was filtered to the first claim line of each claim to eliminate these duplicates. A claim can have multiple claim lines, but every claim will have at least one claim line: the first line. Reducing the data to the first claim line also retains fidelity of the diagnosis, as an inpatient event is bound to an encounter ID. This resulted in a reduction to 9,120,461 records and retained the 44,316 unique patients. This reduction retains the fidelity of the diagnoses while reducing the size of the data by 70.1%.

Step1b – Filter Outpatient/Professional Events

The next data reduction was to remove any outpatient and professional claims while retaining inpatient claims. A readmission event is an inpatient admission. For the purposes of this research, we identify a readmission event as the subsequent admission of a heart failure patient to the same or a different hospital with a specific number of days, with either heart failure or any other condition as the clinical diagnosis. To be qualified as a readmission event, the initial event must be an inpatient event, and the subsequent readmission must be an inpatient event. This significantly reduced the number of candidate patients to 34,990. The majority of data in the database relates to professional and outpatient claims, such as routine checkups and minor outpatient surgery. These records contribute no information around an inpatient admission event for heart failure, they have been filtered from analysis. Similarly, outpatient events are patients who received treatment without being admitted into a hospital; such events are not considered by CMS for a reimbursement penalty, and as such, are also removed.

The resulting dataset represents candidate inpatient records of 208,063 inpatient records. The steps in the following subsection are provided by the CMS model.

CMS Model Criteria

The data exclusion process mirrored as closely as possible the steps in the CMS model of identifying the universe of eligible readmissions, referred to as 'index admissions.' An index admission is an admission in which the subsequent inpatient events are evaluated for 30 days. This process identifies patients who are admitted to a hospital for a heart failure diagnosis, and have to be readmitted within 30 days for any reason, known as all cause readmission. The process for filtering for these admissions is described as follows.

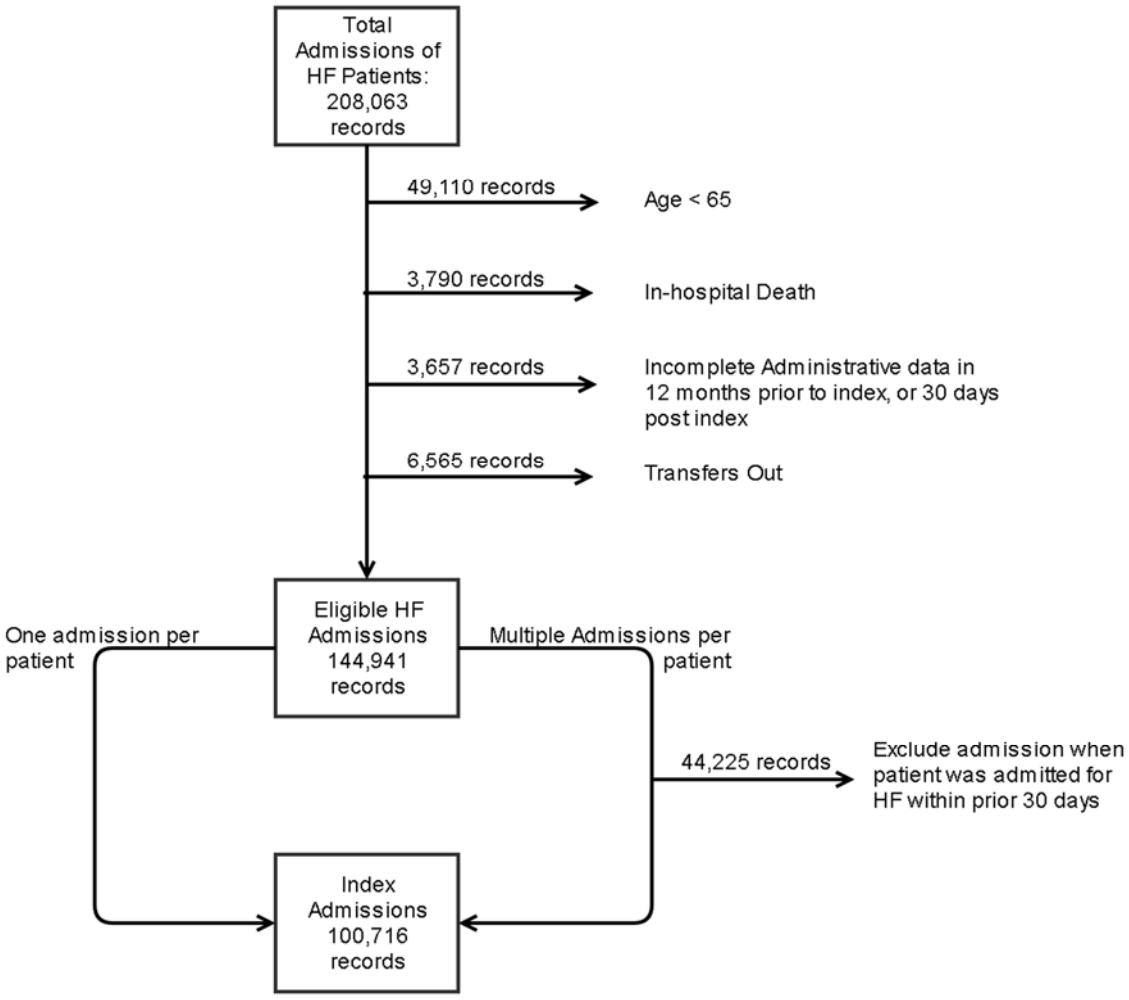


Figure 6. CMS Model Exclusion Criteria.

Step 2

The CMS sanction model removes patients who are under the age of 65. The rationale is that only patients over the age of 65 would use Medicare as their primary insurer. These patient, who use Medicare as their primary insurance mechanism are commonly referred to by CMS, and in this dissertation, as "Medicare primary" patients.

The data was then filtered to remove patients who were under the age of 65 at the time of service. This reduced the number of candidate patients in our data to 25,608 and the number of admission records to 158,953. The reason for this is that individuals who are 65 or older are eligible for Medicare as their primary insurance; CMS penalizes for Medicare readmissions.

Step 3

The CMS model removes any patient who died during service. A patient who died during an inpatient event is ineligible for readmission. Removing these records reduced the candidate patients to 24,968 with 155,163 total inpatient records.

Step 4

The next step in the CMS study is accounting for beneficiaries with 12 full months of Medicare Part A and B. The data does not contain this specific information, but it does contain insurance eligibility history information in a separate table. This information includes pricing and package plans, primary insurers, and effective dates.

Using the 24,968 unique member IDs, the insurance eligibility table was queried using a criterion to filter for any records that were less than 12 months from the start time of service. By applying this filter, the data retains the same fidelity of information pertaining to individuals who were active as payers 12 months prior to the admission, and 30 days after the admission. This reduced the unique members to 24,600 and the number of records with full payer information to 151,506.

Step 5

If a patient is transferred to another acute care facility, the inpatient event becomes confounded and it cannot be known if a readmission has occurred. For example, a patient who was admitted to hospital 5 for a heart failure diagnosis, and was subsequently transferred to hospital 6 that offers specialized cardiovascular care, for the same diagnosis, cannot be known as an admission or readmission because of the transfer. As such, records that had a transfer to another acute care facility have been removed, resulting in 144,941 inpatient records.

Table 11. Summary of Data Extraction and Filtering.

Step	Number of Records	Filter	Result Records	Number of unique Patients
0.	330,967,534	Identify all claims of patients with a history of heart failure	31,240,446	44,316
1a.	31,240,446	Filter to first claim line	9,120,461	44,316
1b.	9,120,461	Filter to inpatient only	208,063	34,990
2.	208,063	Filter to claims where the patient is 65 or older at service start	158,953	25,608
3.	158,953	Filter to remove patients who died during treatment	155,163	24,968
4.	155,163	Filter to remove patients who have incomplete eligibility history (1 year prior, 30 days after)	151,506	24,600
5.	151,506	Filter to remove transfers to other acute care facilities	144,941	18,365
6.	144,941	Filter to remove identical diagnoses on the same day of treatment.	100,716	18,365

Step 6

The final step in the CMS model is to account for heart failure only readmissions that occurred within 30 days, with the rationale being if two heart failure primary cause admissions happened within 30 days, it confounds the readmission date. To control for this, the data was filtered to remove any lines where the member ID, all five diagnosis codes, and the date of services were identical. This resulted in the final dataset of 109,190 records (index admissions) representing 18,365 patients.

Data Transforms

This section describes the transforms that have been applied to the data set of index admissions. These transformations include transforming three columns and creating 15 new columns. The data transformations take one of two forms. First, the data could be transformed from data type into another, such that a string becomes coded as a category. The other kind of transformation is creating a new column based on a calculation of existing columns. This section describes the processes for each of these transformations.

Table 12. Summary of Data Transformations.

Name	Requirement	Process
Age	Calculated Column	Date difference of date of birth and service start date
Diagnosis	Transformed Column	Remove '.' In ICD9s, ie: 404.93 becomes 40493
CC2013_x	Calculated Columns (5)	Join Diagnosis code on CC code where diagnosis = cc
Hccgrouping_x	Calculated Columns (5)	Label group 1 to 25, by CC groups
Sex	Transformed Column	Categorical variable of patient's sex
Medicare	Transformed Column	Categorical variable of Medicare as primary payer
ReadmitDays	Calculated Column	Number of days since service end to service start
All30	Calculated Column	Categorical variable if the patient was readmitted in 30 days
All60	Calculated Column	Categorical variable if the patient was readmitted in 60 days
All90	Calculated Column	Categorical variable if the patient was readmitted in 90 days

Age

The first column that has been added to the dataset is an accurate age at the time of admission. This column is a calculated column by taking the differences of date at which the services were started for the patient and the member's date of birth. The age column represents the member's age at time of service.

Diagnosis

The next transformation that was applied to the dataset is transforming the diagnosis columns to remove the '.' in a diagnosis. For example, the formal specification of hypertensive heart and chronic kidney disease with heart failure and end stage renal disease is '404.93'; this number has been transformed to 40493. This transformation is necessary as the hierarchical condition categories are expressed in this format, and the source diagnosis columns need to be in the same format to match the diagnosis to the condition category.

CC2013_x

To create the condition category columns, five new calculated columns have been added. These hierarchical condition categories (HCCs) are using the CMS 2013 specification (Centers for Medicare and Medicaid Services, 2016). Condition categories (CCs) group together similar diagnoses into a larger category for the purposes of risk adjustment (Pope et al., 2000). HCCs use all diagnoses to predict a total expenditure based on a clinical profile of HCCs. For example, CC 80 groups together all of the 428.x

heart failure diagnoses with 402.xx and 404.xx which are hypertensive disorders with heart failure and kidney failure with heart failure, respectively. Not every ICD9 will belong to a CC, since HCCs adjust for high risk patients. In order to populate the columns created for CCs, the 2014 updated version of HCCs was obtained from CMS and loaded into the MySQL database. From there, an update query was used that populated a value for each CC column corresponding to each diagnosis column, if the diagnosis matched the condition category. For a patient with value of 428.21 as their secondary diagnosis, the corresponding CC2013_CC2 column would have a value of 80. The HCC 2013 specification includes 70 condition categories in 25 groups (Centers for Medicare and Medicaid Services, 2016; NBER, 2016).

Hccgrouping_x

The next calculated column uses a modified version of the 25 groups of condition categories. In a study examining the diagnoses and timing of 30-day readmissions, these groupings were used to better understand the profile of diagnoses of patients (Dharmarajan et al., 2013). There are at least two versions of CCs put forward every year by CMS, and these groupings work as a mechanism to bridge the incremental changes that come with each revision. Since these groupings are based on CCs, they have been calculated using the 2013CC_x columns, for each record that has a CC present. These groupings put together multiple CCs into one umbrella to form categories such as infection, metabolic, cardio, lung, kidney, etc.

Sex/Medicare

The gender column in the data has been transformed from M/F to 1/0, respectively. Similarly, the Medicare as a primary insurer column has been transformed from Y/N to 1/0, respectively. This allows for easier processing, as any statistics platform is going to encode a string variable into a sequential variable. In doing so before analyzing the data, the coding of male/female and Medicare primary is pre-defined and will not change on a per-analysis basis.

ReadmitDays and All30, All60, All90

The data does not contain a readmission column, but it does contain service dates, including service start and service end dates, as well as member IDs, discharge statuses, and diagnoses columns. To calculate readmission, the follow steps were applied: first, the data was sorted by member ID and then by service start date. Next, if the same member had a service start date of a newer time than a service end date, then the diagnosis columns were checked. Using the diagnosis columns, if a member was admitted with a primary diagnosis of any of the inclusion criteria ICD9s found in table 10, then the next inpatient event for that patient was examined. If the patient was readmitted for any cause other than a heart failure, then a value is populated in the 'ReadmitDays' column that takes the difference of the service start date of the readmission and the service end date of the initial admission. This value is then used to populate the 'All30', 'All60', and 'All90' columns. If the 'ReadmitDays' column is less than 31 days, then a value of 1 is coded in the 'All30' column, otherwise it is coded as a 0. If the 'ReadmitDays' column is less than

61 days, then a value of 1 is used for the 'All60', and similar for All91 if the readmission is less than 91 days.

1. Sort data by MemberID
    a. Sort Data by service start Date
2. If service start date < service end date, check diagnosis
3. If diagnosis of record n-1 has a primary diagnosis of heart failure, check record n
4. If record n has any diagnosis other than heart failure, calculate (service end – service start).
    a. Record that value as ReadmitDays
5. If ReadmitDays < 31, then code All30 as 1
    a. If ReadmitDays <61, then code All60 as 1
    b. If ReadmitDays <91, then code All90 as 1

With these transforms, the data mirrors the data used in the CMS model. By the same data fields as the CMS model used, this dissertation is able to replicate the CMS model using the NCSHP data. The following section provides the results of this replication.

CMS Model Replication

This section presents the results of replicating the CMS generalized logit model (GLM) which was described in Chapter 3. These results serve to provide a baseline of prediction accuracy from which this dissertation can improve and expand. The results of this replication show the predictive accuracy of the CMS model using the NCSHP data, and allow for a direct comparison between the model that CMS uses and the predictive ability of alternative modeling techniques.

The CMS model includes 37 variables, 35 of which are CC codes. The remaining two variables are age and gender. The version of the CC codes adopted in the original study is not provided. This replication, and by extension this dissertation, has adopted the 2013 version of the CC codes. Additionally, the original model narrowed the list of CCs down based on clinical feedback, whereas this study is using all 70 CC codes that are available in the data. From the time the CMS model was written to the time this dissertation was written, there have been bi-yearly revisions to the ICD-9 groupings of CC codes to improve their ability to predict and adjust for risk.

In the original paper, the CMS model was trained on a sample of 50% of the data, and validated on the other 50%, by splitting the data into the years 2003 and 2004 (Krumholz et al., 2008). This replication adopts a similar approach. Whereas the original paper broke training and testing by year, this replication randomly samples 50% of the data for a training set, and then validates on the remaining 50%.

The original paper compares the CMS model to the 'gold standard chart data' model. This chart data refers to medical data available on a patient's chart, such as blood pressure, pulse, and vital signs. Chart data for the patients was not available in the NCSHP data. This dissertation omits the chart data comparison of the CMS study. While this limits the completeness of the replication, it does not limit the baseline comparison to advanced analytics. The original model was developed as an alternative to the chart data model, and the original study compared their model to the chart data model. In the original CMS study, they demonstrated that their administrative model performed at

parity and slightly above the gold standard. This dissertation is exploring alternatives to prediction using administrative data, which makes the administrative results of the CMS study of interest, and not the CMS replication of the gold standard.

ROC Curves

This research adopts the ROC curve to compare advanced analytics with the baseline ROC established in this replication. The ROC curve is a graphical visualization that plots true positive rates against false negative rates. The area under the ROC curve is commonly called the c-statistic, and is a standard measure of predictive ability. The ROC is one of the most commonly recognized mechanisms to identify predictive power of all predictive analytics algorithms, including various forms of regression such as (the (Fang, Hu, Li, & Tsai, n.d.; Guzella & Caminhas, 2009; Stelfox et al., 2011). A ROC curve plots the position of true positives against false negatives at any given point based on a confusion matrix.

Because an ROC curve uses true positives and false negatives, a confusion matrix is used and can derive additional statistics. A basic confusion matrix is a 2x2 table that compares predicted outcomes against actual outcomes, as referenced in table 13.

Table 13. An Example of a Confusion Matrix.

		Predicted Conditions	
		Predicted Positive	Predicted Negative
True Conditions	Condition Positive	True Positive	False Negative (Type II Error)
	Condition Negative	False Positive (Type I Error)	True Negative

Several additional statistics that can be derived from a confusion matrix include accuracy, positive predicted value (precision), negative predicted value, sensitive (recall), false positive rates (fall-out), and true negative rate (specificity). Accuracy calculates the true positives and negatives against the total population. Precision calculates the number of true positives against the tested outcome positives; negative predicted value does the same with true negatives against test outcome negatives. Sensitivity measures the sum of true positives to condition positives. False positive rates calculate the sum of false positives to the sum of condition negatives. Lastly, specificity calculates the sum of true negatives to the sum of condition negatives. Figure 7 references common statistics derived from a confusion matrix.

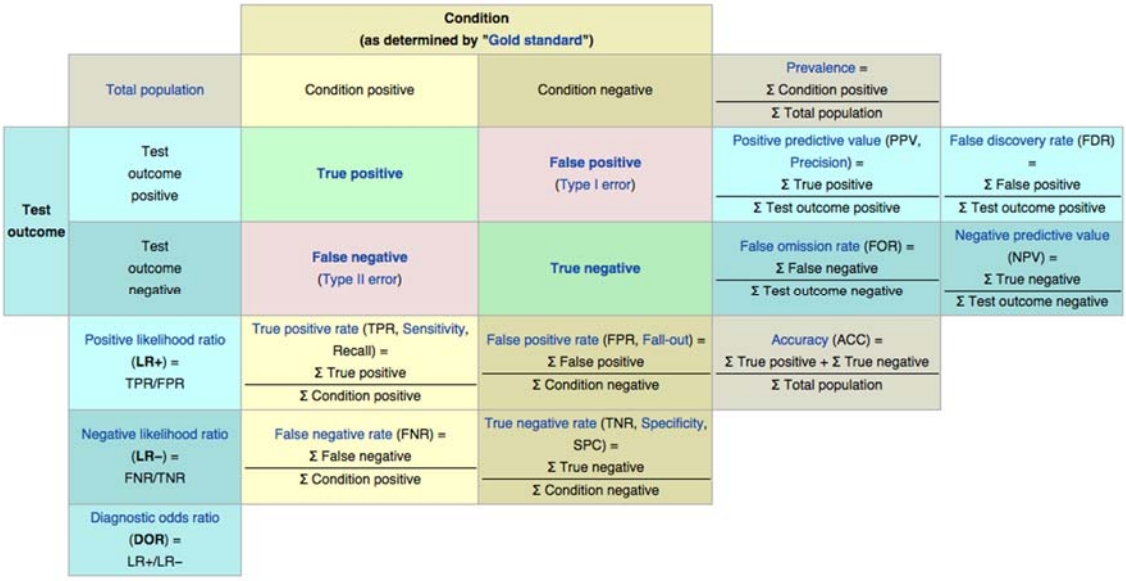


Figure 7. Sensitivity and Specificity Outcomes.

For this dissertation, the metrics of sensitivity, specificity, accuracy, positive predicted value, and negative predicted value will be examined. Sensitivity and specificity are the metrics by which the ROC curves are visualized. Additionally, sensitivity and specificity provide insight in terms of a model's predictive ability's error rates. Because of the nature of the data, positive predicted value and negative predicted value are also reported, with a positive prediction representing a predicted readmission, and a negative prediction representing no readmission.

CMS Model Performance

The original model was assessed with six summary statistics. These summary statistics include: over-fitting indices, percent of variance explained (R^2), predictive ability, area under receiver operating characteristic curve, distribution of residuals, and model χ^2. This replication reports five of these summary statistics, omitting the over-fitting indices. These values are reported in table 14.

The R^2, which provides statistical explanation of each of the variables as it contributes to statistical significance. Accuracy is measured in terms of the predicted to the actual outcome. The ROC statistic provides the area under the ROC curve. Pearson residual fall describes the distance of the residuals from the prediction. Lastly, the Wald Chi-square provides a test for categorical predictors.

Table 14. CMS Model Replication Performance.

		R^2	*Accuracy*	*ROC*	*(Pearson Residual Fall%)*				*Model* χ^2
			(lowest decile, highest decile		<-2	[-2,0)	[0,2)	[2+	[Number of Covariates]
Model									
50% Training	N=50,358	0.007	0.15-0.43	0.618	0%	98.44%	0.61%	0.94%	3403 (331)
50% Validation	N=50,358	0.007	0.15-0.50	0.611	0%	98.48%	0.42%	1.00%	3199 (331)
100% of Data	N=100,716	0.008	0.62-0.98	0.613	0%	98.46%	0.48%	1.04%	6590 (331)

The summary statistics in the replication were close to matching the original study. Of interest is the original study achieved an ROC of .610, a number almost identical with the replication. The R^2 values are lower than the original, but this is attributed to the number of CC codes doubling. The accuracy of the model appears to be stable as well, with the original study achieving a 0.15 to 0.37 predictive ability, and the replication achieving very similar, albeit slightly higher results. This can be attributed to a smaller sample size. Lastly, the residuals were within expectations as well, with the majority falling within -2 or +2 of the mean. Overall, replicating the CMS GLM model generated very comparable results, with the key metric of the ROC being nearly identical.

Where the replication diverges is in the significance of the CC codes. Whereas the original study found over 20 CC codes statistically significant, this replication only found five total, with CC80 overlapping. Table 15 references the training set, which found CC174 statistically significant; CC174 is defined by major organ transplant procedures.

CC80 combines heart failure conditions together. In the validation set, we find CC80 together with CC92, specified heart arrhythmias, as well as CC131, renal failure. Lastly, on the full data set, CC130, dialysis, as well as CC79, Cardio-respiratory failure and shock, show a statistical significance. These results provide insight as to the nature of conditions that are present in patients in this data set. While none of these conditions are a new finding, this replication does provide further empirical evidence to the known body of medical knowledge.

Table 15. Significant Variables from Training Results Replication. (50%).

Variable	Estimate	Standard Error	Z value	P-value
Age	0.044	0.004	8.790	<0.001
CC174 (Major Organ Transplants)	3.880	1.337	2.902	0.004
CC80 (Congestive Heart Failure)	3.225	.714	4.518	<0.001

Table 16. Significant Variables from Validation Results Replication. (50%).

Variable	Estimate	Standard Error	Z value	P-value
Age	0.030	0.004	6.102	<0.001
CC131(Renal failure)	-1.946	0.808	-2.408	0.016
CC80 (Congestive Heart Failure)	-2.063	0.786	-2.624	0.008
CC92 (Specified Heart Arrhythmias)	-1.814	0.794	-2.284	0.022

Table 17. Significant Variables from Replication. (100%).

Variable	Estimate	Standard Error	Z value	P-value
Age	-.108	0.003	10.568	<0.001
CC130(Dialysis Status)	2.665	0.930	2.865	0.004
CC174(Major Organ Transplants)	4.091	1.286	3.180	0.001
CC79(Cardio-Respiratory Failure/Shock)	1.517	0.752	2.017	0.044
CC80 (Congestive Heart Failure)	3.880	0.711	5.461	<0.001

In addition to the statistics reported in the CMS model, this dissertation also examines the characteristics of the ROC curves for each result. Sensitivity, specificity, accuracy, positive predicted value (PPV), and negative predicted value (NPV) are reported. These characteristics can be found in table 18.

Table 18. ROC Coordinates for CMS Replication.

Model	Predictive Accuracy Measures					
	Sensitivity	Specificity	Accuracy	PPV	NPV	AUROC
50% Train	0.577	0.623	0.628	0.024	0.989	0.618
50% Validate	0.617	0.572	0.573	0.022	0.989	0.611
100% Total Data	0.519	0.681	0.678	0.025	0.989	0.612

As reported earlier, the ROC in this replication is very close to that of the original paper. The numbers from table 18 are coordinates at the 'best' position on the curve - their highest point. To further expand on this interpretation, at best with the 50% training set, the model was 62.8% accurate with 2.4% of the data representing a predicted positive readmission and 98.9% of the data representing a no-readmit.

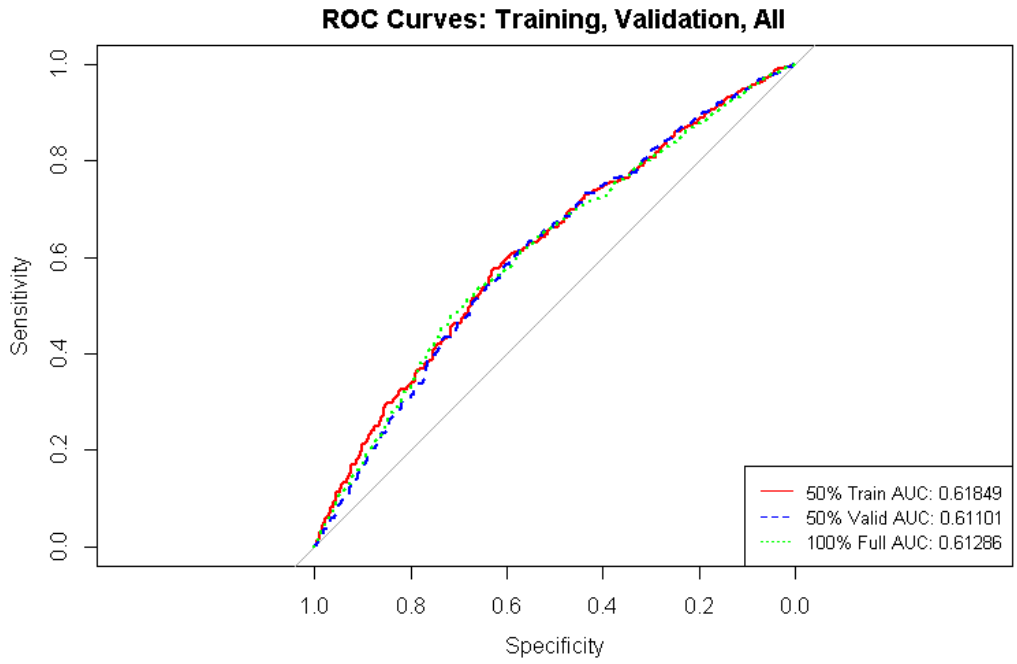


Figure 8. ROC Curves for Baseline Comparison .

Figure 8 provides the ROC curves for each of the three models that were assessed in this chapter. The 50% training set performed the best with a 0.618 area under the curve. The 100% data set also performed in between both, as it should. To improve and expand, the ROC must be demonstrated above these current lines, as shown in the next chapter.

This chapter concludes with reapplying the CMS GLM model on new training intervals. These intervals are to account for model overfitting. These intervals are 50%, 60%, 70%, 75%, and 80%. These levels have been chosen because overfitting of prediction models tends to occur between 70 and 80% (Wongravee, Lloyd, Silwood, Grootveld, & Brereton, 2010). Figure 9 illustrates the ROC curves for each of these

intervals that will serve as a baseline comparison for the advanced analytics, and table 19 provides the relevant ROC coordinates for those curves.

As the training interval increases, the ROC curve also increases. It is demonstrated here that at 80% training, the model has not yet overfit. Referencing the prior model using 100% of the data, the model was overfit with a decrease in ROC to 0.612. Each of these levels will be used in the next chapter that examines alternatives to the GLM model.

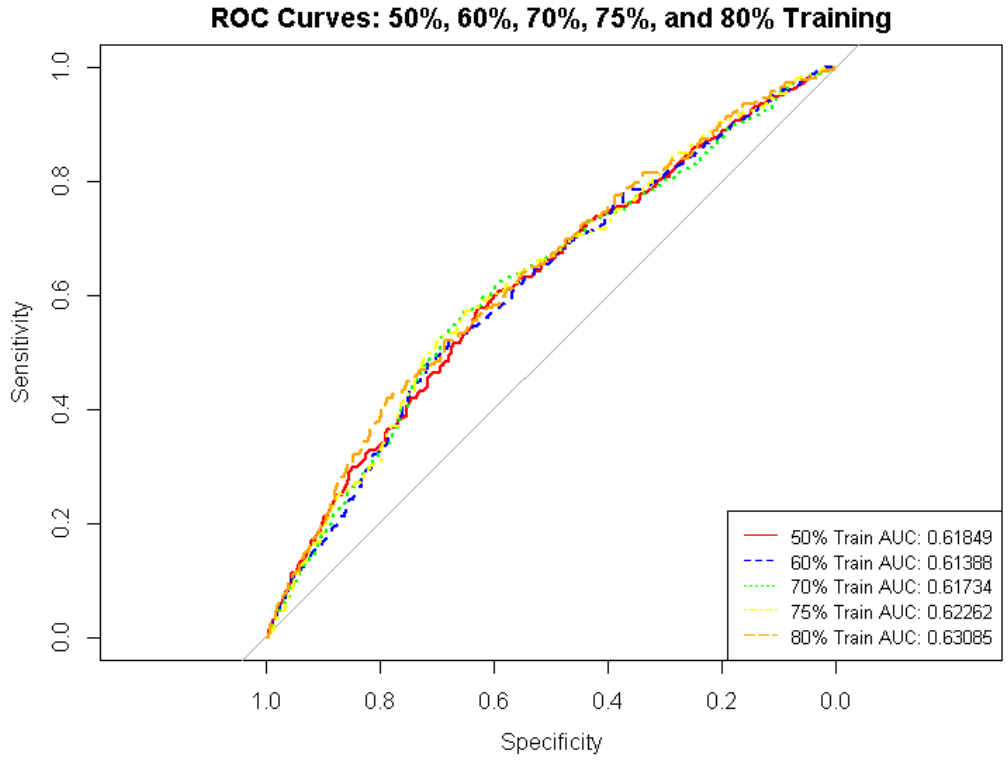


Figure 9. ROC Curves for 50, 60, 70, 75, and 80% training.

Table 19. ROC Coordinates for 50%, 60%, 70%, 75%, and 80% Replication Training.

Model	Predictive Accuracy Measures					
	Sensitivity	Specificity	Accuracy	PPV	NPV	AUROC
50% Training	0.577	0.623	0.628	0.024	0.989	0.618
60% Training	0.521	0.679	0.677	0.025	0.989	0.613
70% Training	0.571	0.658	0.656	0.026	0.990	0.617
75% Training	0.571	0.650	0.689	0.026	0.989	0.623
80% Training	0.521	0.689	0.686	0.022	0.991	0.630

CHAPTER V

ANALYTICS RESULTS

This chapter presents the results of analytics to model 30 day all cause heart failure readmissions. These analytics include conditional inference decision trees, self-organizing maps, the naïve Bayes classifier, and artificial neural networks. Each result of each analytic is compared using the area under the receiver operating curve (AUROC) c-statistic described in Chapter IV. This statistic will uniformly be referred to as ROC throughout the results.

There are four subsections in this chapter, one for each analytic being utilized. Each result will be presented uniformly as follows. First, the advantages in terms of the results will be discussed in the introduction of each analytic result. Next, the overall performance of the analytic, as measured by the ROC will be presented; ROC results will be for five training levels (50%, 60%, 70%, 75%, and 80%). After presenting the ROCs, a closer examination of the results of each analytic will be presented for the training level that performed the best. This chapter concludes with a summary comparison of each analytic to the baseline ROC.

For assessing the performance of each analytic, the sections will first present the ROC results. For the best training/validation set for each analytic, six statistics are presented: sensitivity, specificity, accuracy, positive predicted value, negative predicted value, and ROC. For reference, sensitivity measures the sum of true positive values

divided by the predicted positive values; specificity measures the true negative values divided by the predicted negative values. Accuracy is calculated by adding the sum of the true positive and true negative values and dividing by the total population. Positive predicted values measure the ratio of true positive to predicted positive; negative predicted values measure the ratio of true negative to predicted negative. Lastly, ROC is the percentage of area under the receiver operating curve, with ROC representing the ratio of true positive to false positive rates.

Conditional Inference Decision Trees Results

This section presents the results of conditional inference trees as an alternative to modeling readmission. Conditional inference trees are decision trees that create decision splits based on statistical significance. The splits in the tree function as decision heuristics in a manner that can classify a variable. For this specific result, the variable being classified is thirty day, all cause readmission. The classifiers are the five diagnosis columns present in the data, in addition to age and sex.

To account for overfitting the models, the five training levels have been tested and their ROC curves reported as presented in table 24. This section provides an in-depth report of the model which presented the best predictive power without overfitting, the 75% training set.

In order for a new node, or decision, to be generated to classify readmission, a variable had to provide a p-value of less than 0.05. These trees were generated with the

'party' library (Hothorn et al, 2006). The party library uses conditional inference trees to classify variables, and was chosen over the rpart library in that the ctree method of the party library uses a significance test in order to select variables instead of selecting variables that maximize information gain. Whereas the rpart library excels at generating multiple nodes, the ctree algorithm is resistant to generating node unless a statistical significance is present. In so doing, the ctree method is resilient to overfitting in a manner that the rpart method is not.

The results of the tree are discussed at length using a depth-first narrative of the path that classifies the most readmissions. This narrative allows the results of the decision tree to be interpreted by assessing the decisions the tree made, and how those decisions relate to understanding the profile of a patient who is at risk of a heart failure readmission.

For modeling the data using decision trees, the analysis kept the same variables used in the CMS model, predicting a binary response of 1 if the patient was readmitted. The same dataset of 100,719 admissions was used for both the decision trees as well as the CMS logit model. The next section discusses the results of the 75% training level, which performed with an ROC of 0.819 without overfitting, an improvement of 19.7% over the logit model.

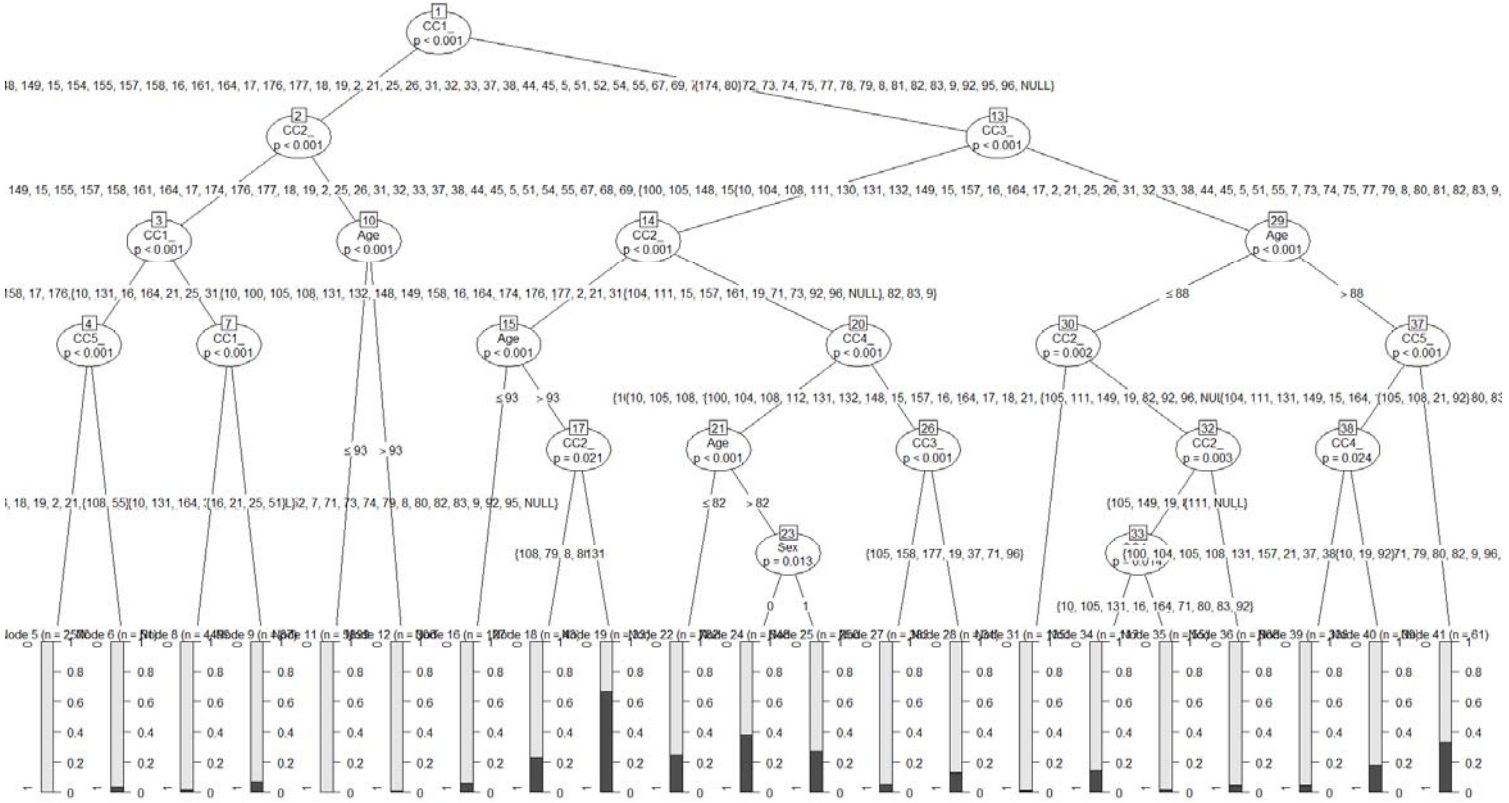


Figure 10. Conditional Inference Tree. 75% Training.

As an overview of the decision path, the algorithm provided the following result to classify the most readmissions:

1. Is the primary diagnosis a heart failure or transplant? (Node 1 to 13)
2. What are tertiary reasons for the admission? (Node 13 to 14)
3. What are the secondary reasons for the admission? (Node 14 to 15)
4. Is the patient over 93? (Node 15 to 17)
5. Does the patient have a renal failure? (Node 17 to 19)

This particular expands upon the CMS logit model in that it provides multiple criteria for each CC code recursively. Where the CMS logit model provides statistical significance if a person has been coded with a particular CC, this model can account for that CC multiple times at each node in the decision tree, which provides a more holistic view of a profile of a patient at risk for a heart failure readmission.

The first decision the algorithm provides is the patient's primary diagnosis. The algorithm splits into a variety of different decisions, with no uniform decision at each sub-level. To better explain these results, an examination of the path that predicts the most readmissions is discussed below.

Table 20. Decision Tree - First Decision Split.

CC Number	Description
80	Congestive Heart Failure
174	Major Organ Transplant Status

The decision tree's first split categorizes a variety of CC codes into the split to node 2, and then only 80 and 174 to node 13. This decision is explained as the algorithm examining the primary diagnosis. If the primary diagnosis was a congestive heart failure

or major organ transplant, then it advances to Node 13, which examines the tertiary diagnosis.

This is the second decision that is provided by the algorithm. A patient who has a third CC code matching any of the codes in table 20 will advance the path to the terminal node (19) that contains the most readmissions. It is of note that these particular CC codes come from a variety of disorders and ICD-9 diagnoses, as opposed to the commonly excepted pulmonary or vascular diagnoses.

Table 21. Decision Tree - Second Decision Split.

CC Number	Description
18	Diabetes with Ophthalmologic or Unspecified Manifestation
19	Diabetes without Complication
37	Bone/Joint/Muscle Infections/Necrosis
54	Schizophrenia
67	Quadriplegia, Other Extensive Paralysis
68	Paraplegia
69	Spinal Cord Disorders/Injuries
71	Polyneuropathy
96	Ischemic or Unspecified Stroke
100	Hemiplegia/Hemiparesis
105	Vascular Disease
148	Decubitus Ulcer of Skin
158	Hip Fracture/Dislocation
161	Traumatic Amputation
177	Amputation Status, Lower Limb, Amputation Complications

The next decision that was determined by the algorithm was to examine the secondary diagnosis. If a patient was classified to have any of the diagnoses found in table 21, then the tree moves to the next decision node, found in table 22.

Table 22. Decision Tree - Third Decision Split.

CC Number	Description
2	Septicemia/Shock
7	Metastatic Cancer and Acute Leukemia
8	Lung, Upper Digestive Tract, and Other Severe Cancer
9	Lymphatic, Head and Neck, Brain, and Other Major Cancers
10	Breast, Prostate, Colorectal and Other Cancers and Tumors
16	Diabetes with Neurologic or Other Specified Manifestations
21	Protein-Calorie Malnutrition
31	Intestinal Obstruction/Perforation
32	Pancreatic Disease
33	Inflammatory Bowel Disease
44	Severe Hematological Disorders
45	Disorders of Immunity
51	Drug/Alcohol Psychosis
55	Major Depressive, Bipolar, and Paranoid Disorders
67	Quadriplegia, Other Extensive Paralysis
69	Spinal Cord Disorders/Injuries
75	Coma, Brain Compression/Anoxic Damage
79	Cardio-Respiratory Failure and Shock
80	Congestive Heart Failure
81	Acute Myocardial Infarction
82	Unstable Angina and Other Acute Ischemic Heart Disease
83	Angina Pectoris/Old Myocardial Infraction
100	Hemiplegia/Hemiparesis
105	Vascular Disease
108	Chronic Obstructive Pulmonary Disease
131	Renal Failure
132	Nephritis
148	Decubitus Ulcer of Skin
149	Chronic Ulcer of Skin, Except Decubitus
158	Hip Fracture/Dislocation
164	Major Complications of Medical Care
174	Major Organ Transplant
176	Artificial Openings for Feeding or Elimination

177	Amputation Status, Lower Limb/Amputation Complications

The next decision, at node 15, examined the secondary diagnoses to create a split on a patient's age. For this dataset, if the patient was over the age of 93, the decision tree advanced to node 17. For patients over the age of 93, one more separation of the secondary diagnoses was made, with a terminal node at 19. If a patient had a secondary diagnosis of CC131, renal failure, then the decision tree terminated at leaf node 19, with 68% of the observations that fit this profile of a patient falling into a readmission category.

This section has examined the decision heuristics generated to classify readmissions that account for the majority of readmissions in the data. There are multiple other paths, but none present as many readmits in their terminal/leaf node. The next section provides the decision tree outcomes for each training level.

Each of these decision heuristics is able to model readmission in a way that the CMS logit model did not. In so doing, the classification decision tree is able to expand upon current knowledge of understanding variance in heart failure readmissions.

Table 23. Decision Tree Classification Profile for Heart Failure Readmissions.

Node number	Decision to Advance Tree	Interpretation
1	CC1 = {10, 100, 101, 104, 105, 107, 108, 111, 112, 130, 131, 132, 148, 149, 15, 154, 155, 157, 158, 16, 161, 164, 17, 176, 177, 18, 19, 2, 21, 25, 26, 31, 32, 33, 37, 38, 44, 45, 5, 51, 52, 54, 55, 67, 69, 7, 70, 71, 72, 73, 74, 75, 77, 78, 79, 8, 81, 82, 83, 9, 92, 95, 96} to Node 2 1) CC1 == {174, 80} to Node 13	If the patient has a primary diagnosis of 80 or 174, then examine node 13. Otherwise examine node 2: What is the primary diagnosis?
13	CC3_ == {100, 105, 148, 158, 161, 177, 18, 19, 37, 54, 67, 68, 69, 71, 96} to Node 14	What are the comorbidities of a patient with a primary diagnosis of 174 or 80?
14	CC2_ == {10, 100, 105, 108, 131, 132, 148, 149, 158, 16, 164, 174, 176, 177, 2, 21, 31, 32, 33, 44, 45, 51, 55, 67, 69, 7, 75, 79, 8, 80, 81, 82, 83, 9} to Node 15	What is the secondary diagnosis of the patient that has any of the attributes found in node 1 and 13?
15	Age <= 93 to Node 16 Age > 93 to Node 17	The patient must be over the age of 93.
17	CC2_ == {131} to Node 19	The patient has a comorbidity of renal failure.
19	Terminal Node	A patient who meets this classification of attributes can be more accurately predicted for readmission.

Decision Tree Graphs for each Training Level

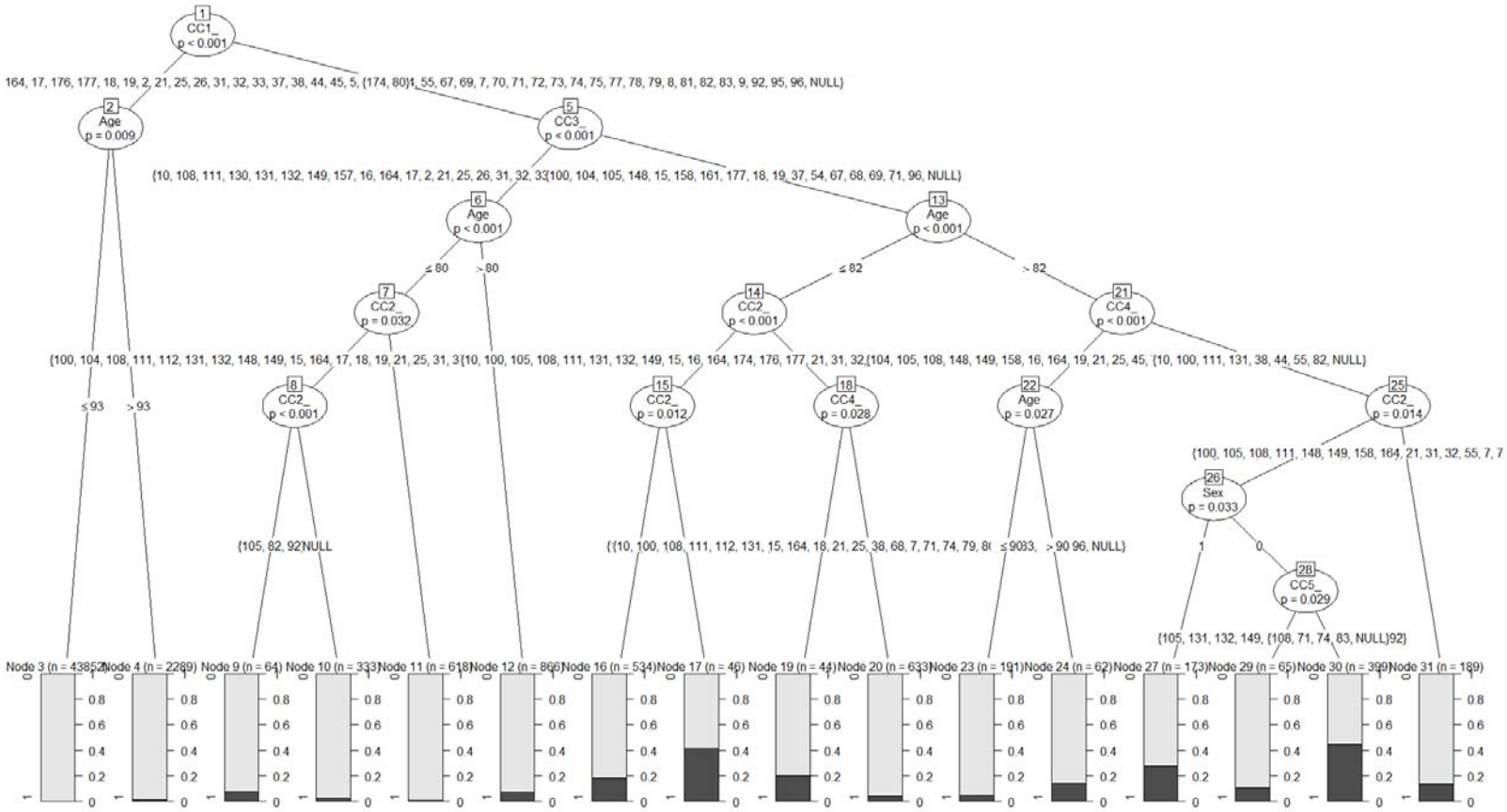


Figure 11. Conditional Inference Tree. 50% Training.

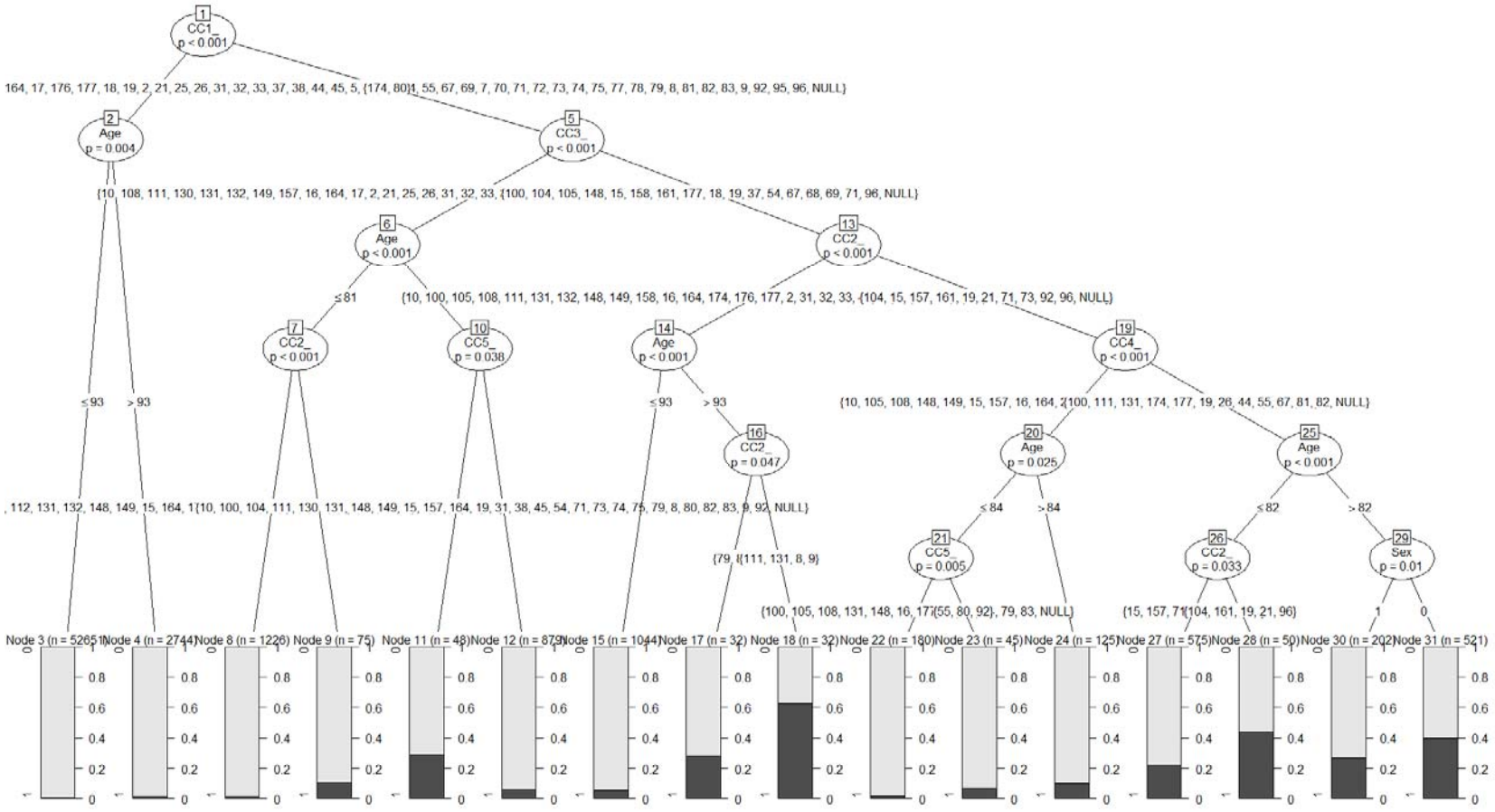


Figure 12. Conditional Inference Tree. 60% Training.

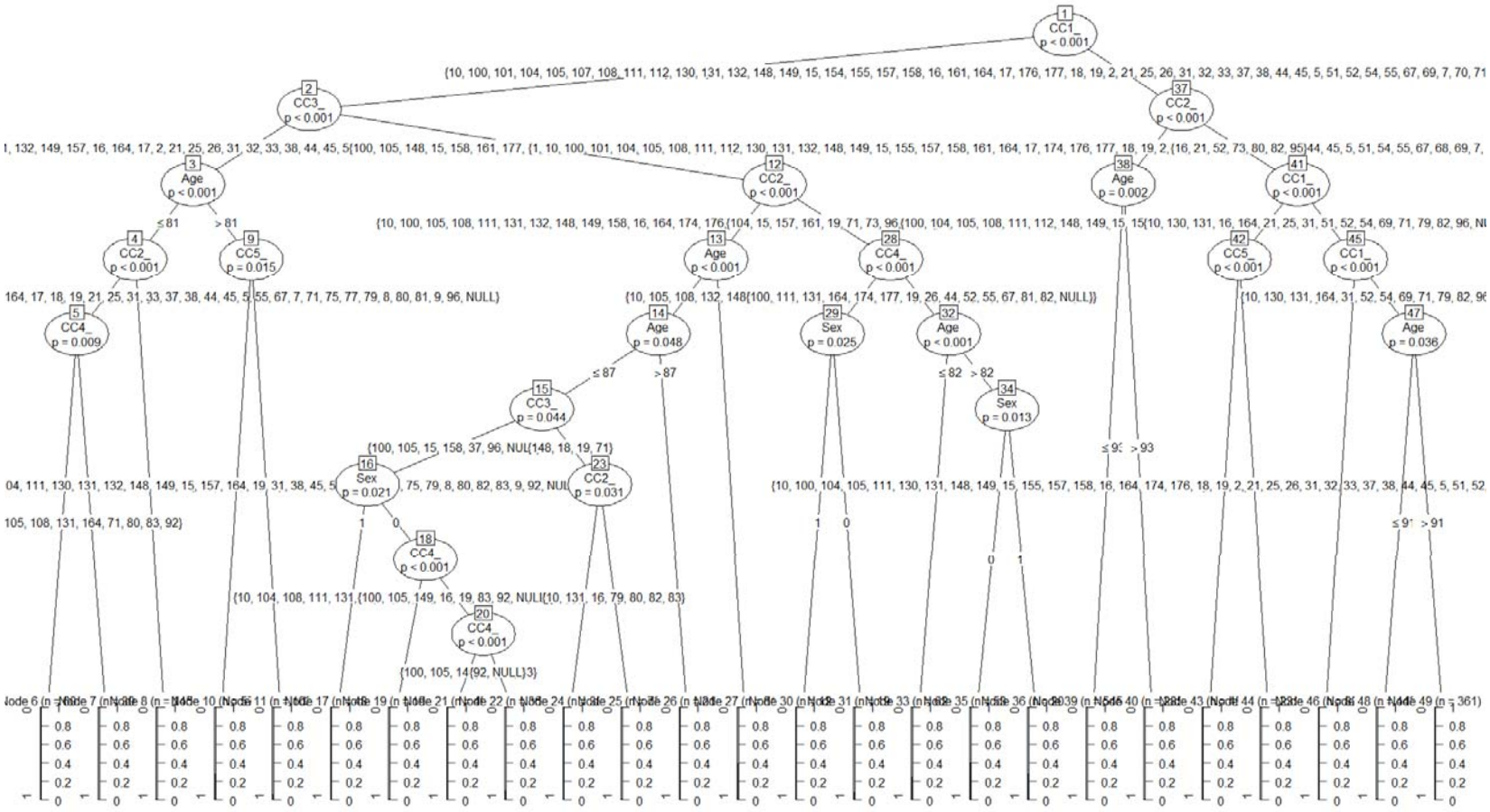


Figure 13. Conditional Inference Tree. 70% Training.

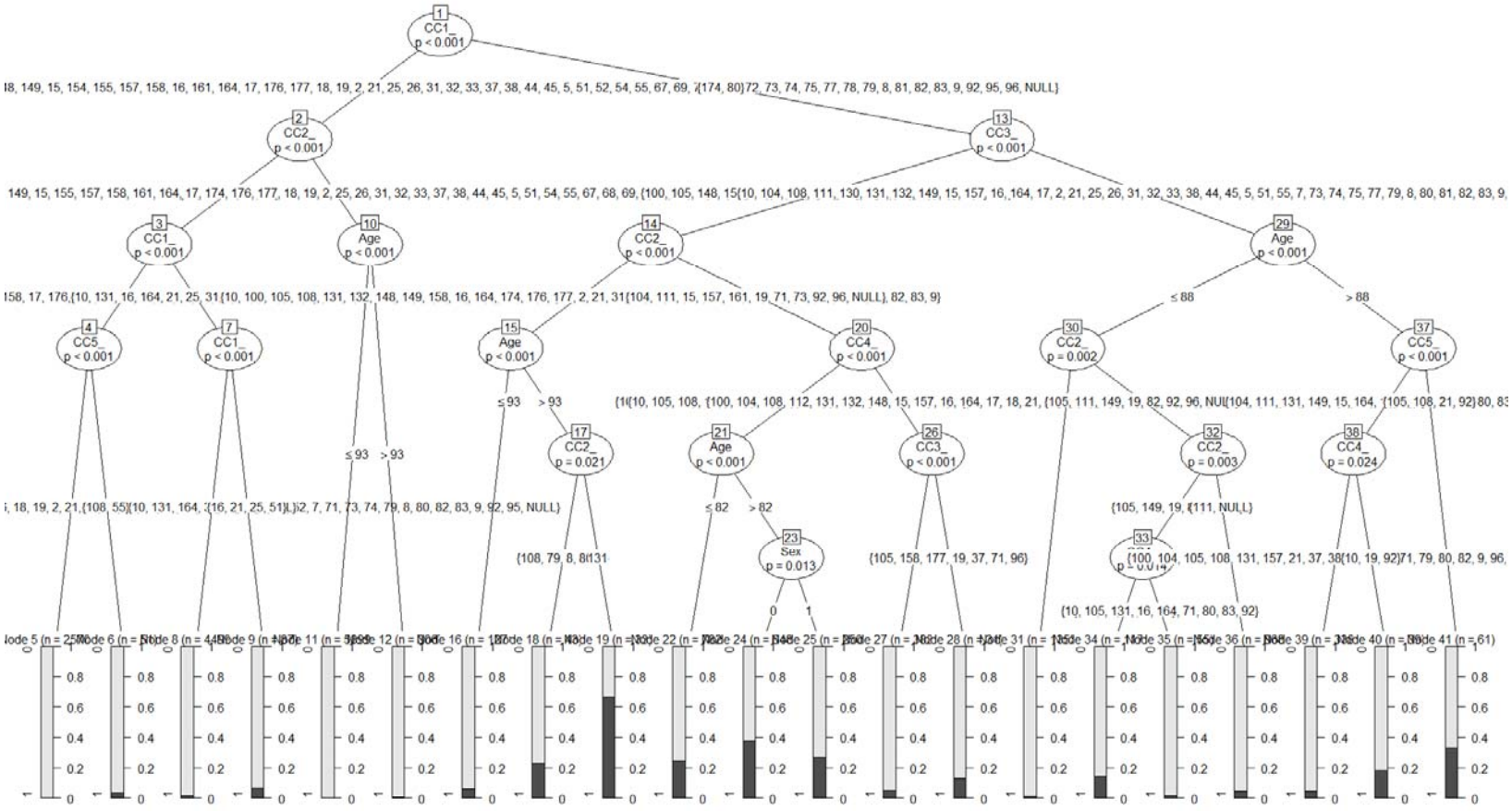


Figure 14. Conditional Inference Tree. 75% Training.

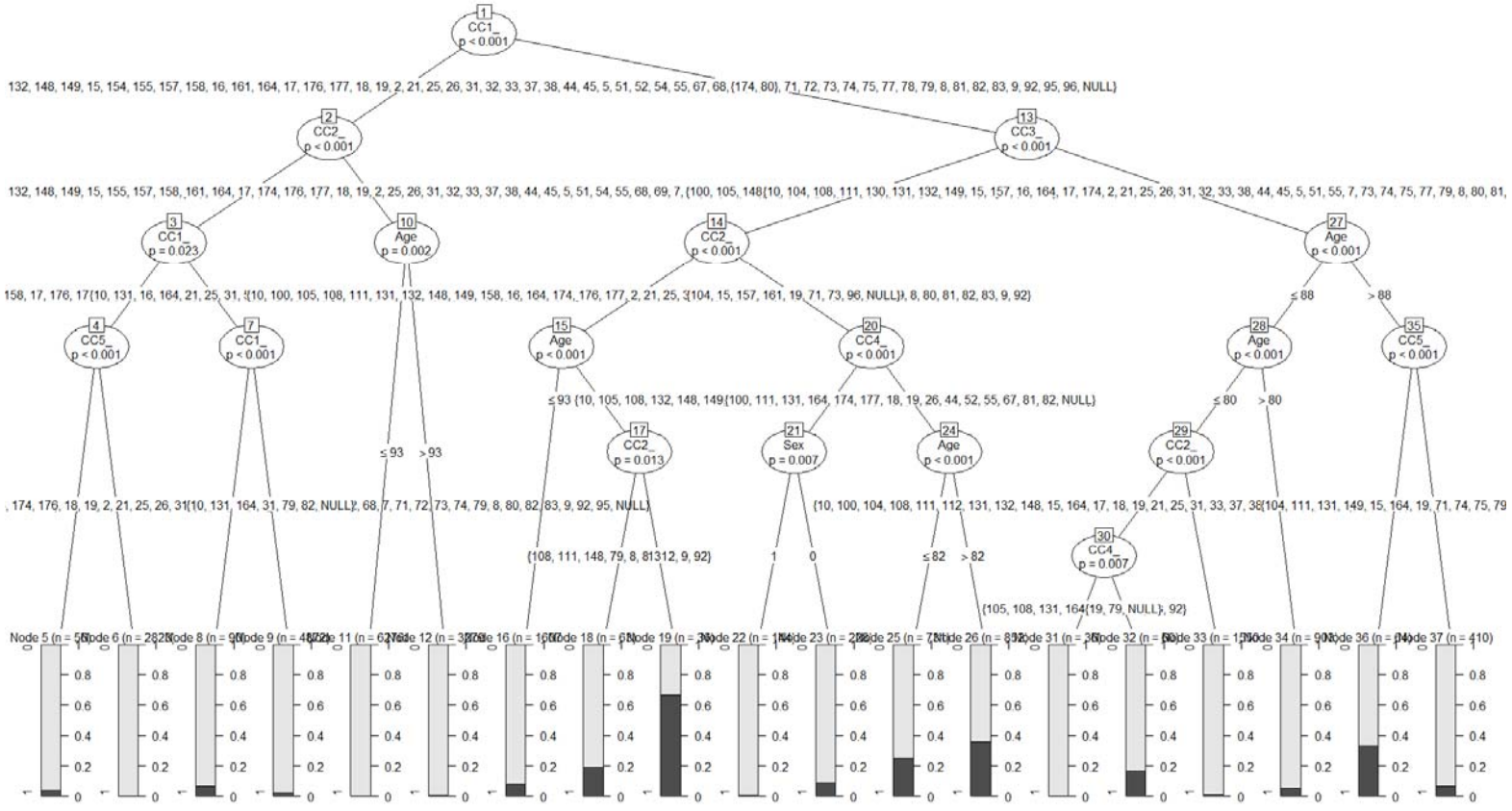


Figure 15. Conditional Inference Tree. 80% Training.

Decision Tree Modeling Validation

The decision tree and CMS logit model ran with five different training and validation levels. The results presented here discuss the 75% training/25% validation model. ROC results were used to determine the best model without overfitting the data. sensitivity, specificity, accuracy, positive predicted value (PPV), and negative predicted value (NPV) are reported.

Table 24. ROC Performance for Decision Trees.

	30Day Readmission	
Train%/Test%	Logit Model ROC	Decision Tree ROC
0%/50%	0.618	0.795
60%/40%	0.613	0.793
70%/30%	0.617	0.817
75%/25%	**0.622**	**0.819**
80%/20%	0.630	0.808

Across each metric, the decision tree provides better performance than the CMS model. It should be noted that the positive predicted values and the negative predicted values are skewed, due to the number of readmits in comparison to the size of the data. Additionally, the CMS model's best prediction rate from the source paper is 0.610. Similar results were generated using the data for this paper, with the best ROC for the CMS model at 0.623. Sensitivity and specificity both exceeded the CMS model by a large margin. The large gains in sensitivity and specificity show a sharp increase on the y-axis of the ROC plot, as well as a steady gain on the x-axis of the ROC plot, covering more area under the curve than the CMS model.

Table 25. Performance Comparison of Decision Trees to CMS Logit Model.

Model	Predictive Accuracy Measures					
	Sensitivity	Specificity	Accuracy	PPV	NPV	AUROC
Baseline	0.571	0.650	0.689	0.026	0.989	0.623
Dtree	0.720	0.864	0.862	0.077	0.995	0.819

The result's specificity presents a higher percentage than sensitivity, which leads to the sharp increase on the ROC plot. This means that the analytic performed better predicting negative values than positive values, meaning that the model was more able to identify non-readmits than readmits. The sensitivity identifies the positive values, which are the readmits, and outperforms the CMS logit model. As evidenced from the plots that follow, the model's specificity lifted the curve above the CMS model at all training levels, with a decline at 80%.

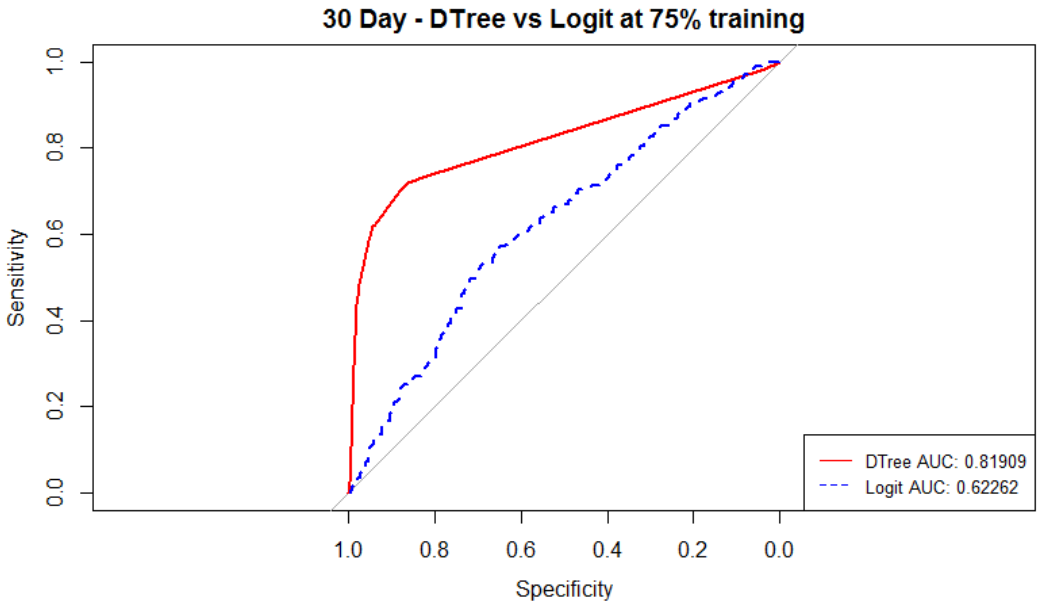


Figure 16. Decision Tree ROC for 75% Training.

Decision Tree ROC Curves – All Training Levels

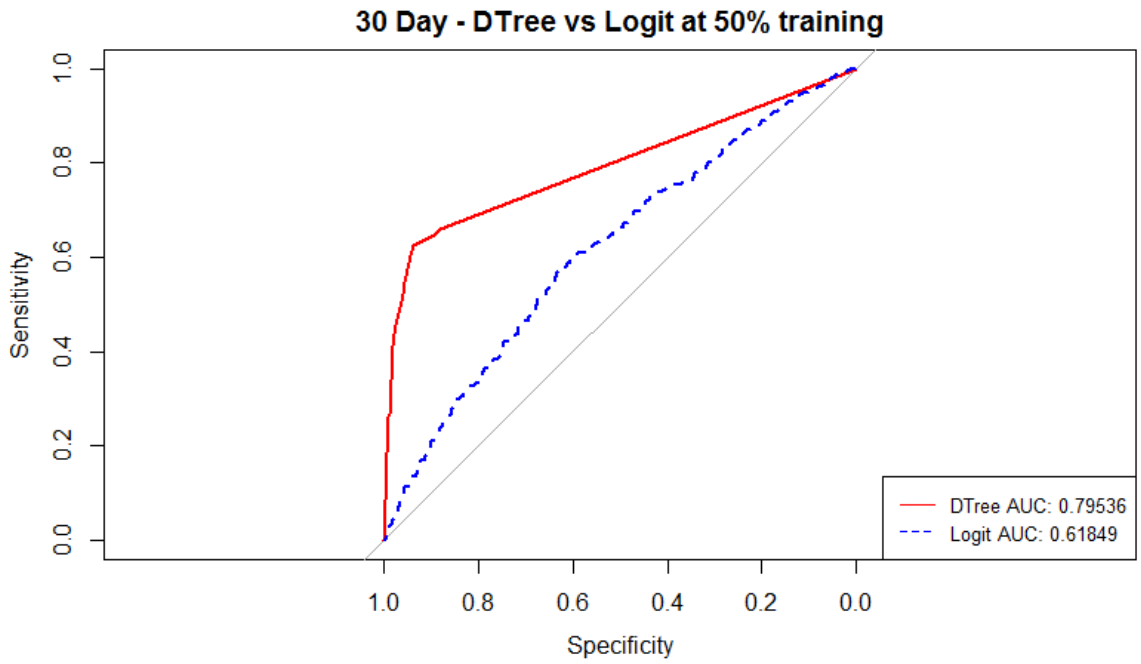


Figure 17. Decision Tree ROC for 50% Training.

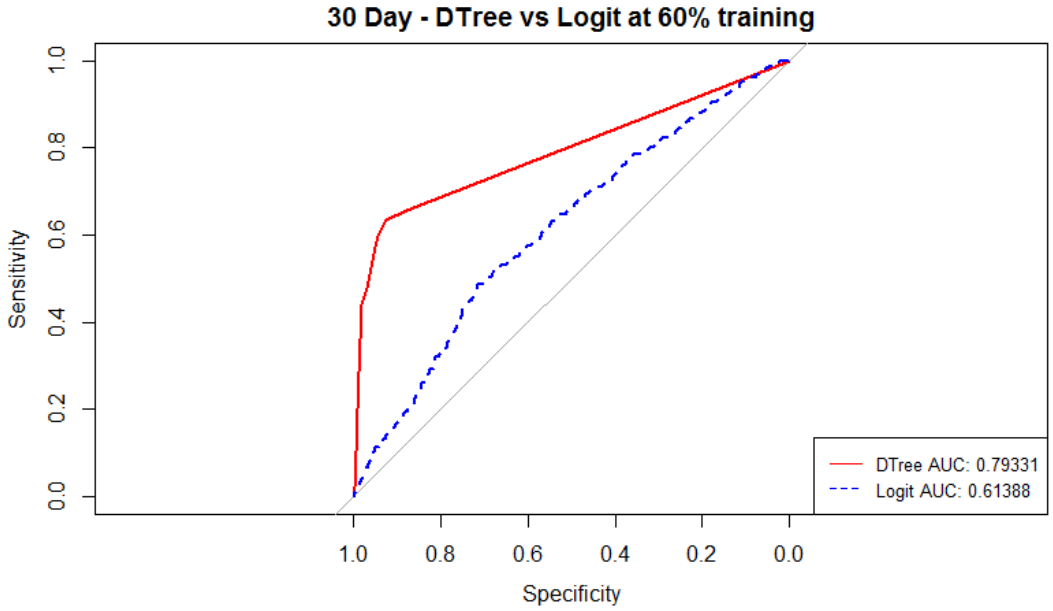


Figure 18. Decision Tree ROC for 60% Training.

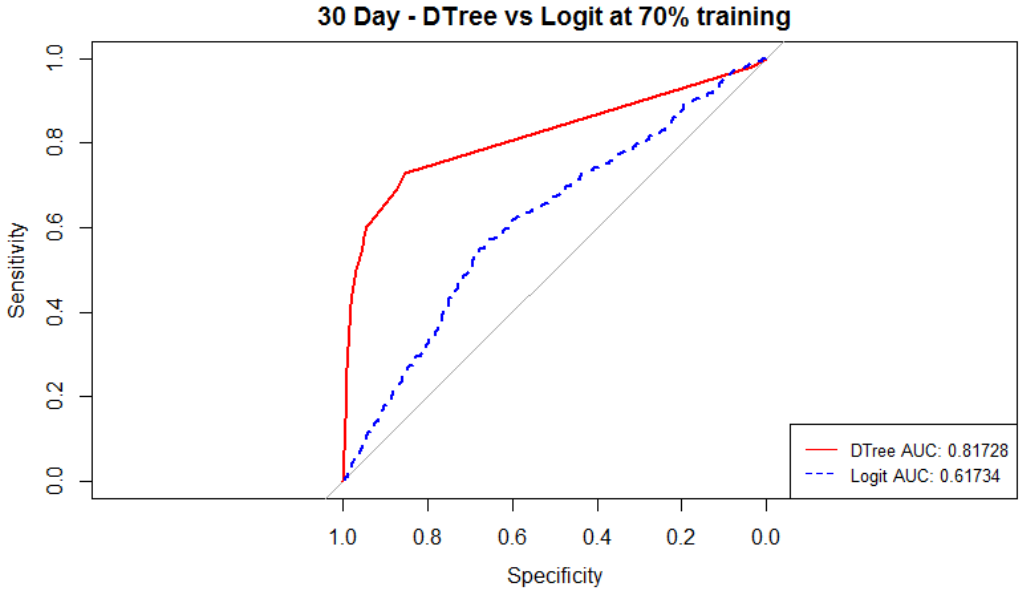


Figure 19. Decision Tree ROC for 70% Training.

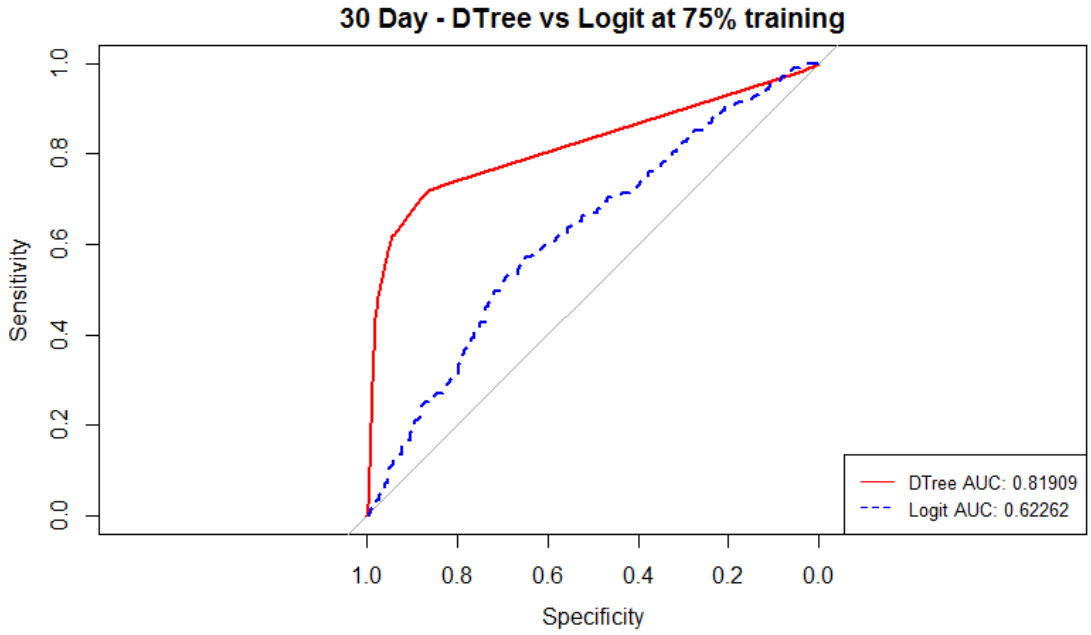


Figure 20. Decision Tree ROC for 75% Training.

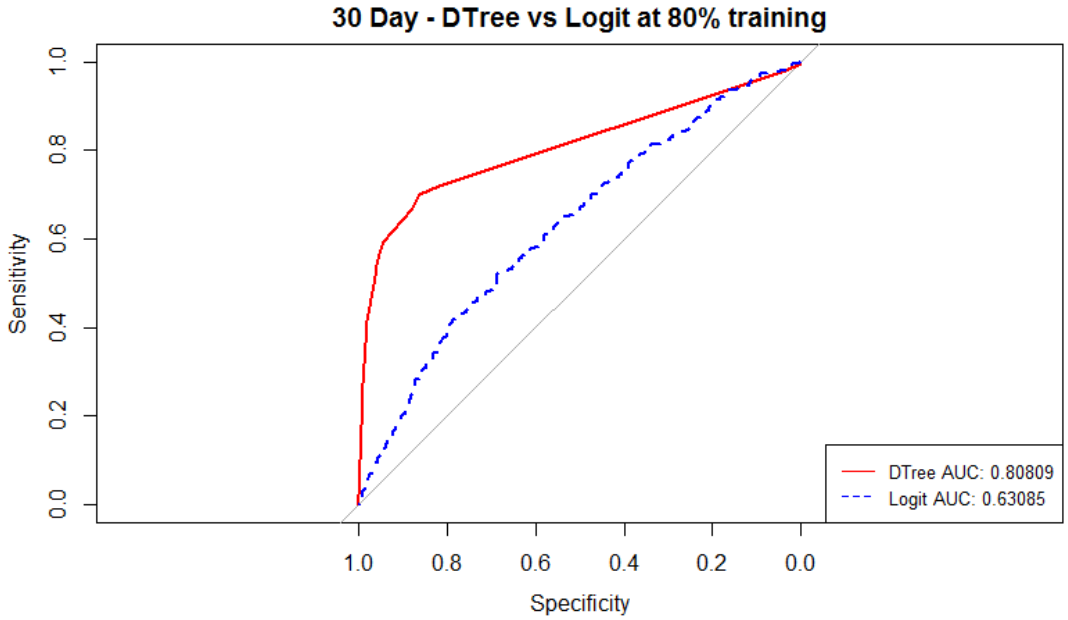


Figure 21. Decision Tree ROC for 80% Training.

Artificial Neural Network Results

The next analytic that is examined is the usage of artificial neural networks to predict all cause heart failure readmissions. The nnet package was used to create the neural network predictions (Venables and Ripley, 2002). Artificial neural networks (ANN) have the advantage of running multiple functions in parallel in order to determine an optimal solution. In doing so, ANN has the advantage of prediction ability, but to the detriment of a black box approach in which it generates *n*-number of functions that comprise the network.

To generate the ANN results, a network of one hidden layer with ten hidden neurons was specified. The number of hidden layers was chosen because of one dependent variable, and the number of hidden neurons was chosen for the number of independent variables plus dependent variables. The *I* nodes represent the initial nodes of the input variables, and the *H* nodes represent the hidden nodes from the hidden layer. The *B* node represents the bias associated with each hidden node; the bias node functions in the same manner as an intercept in a regression model, except the bias is applied to all functions processed in the ANN. The *O* layer represents the output node. The coefficients are referenced below.

The best result of ANN came from the 50% training/50% validation datasets. The model presented a difference of 0.003 predictive ability when increasing the training from 50% to 60%, but then overfit at 70%. These results are most likely attributed to model specification of ten neurons in the hidden layer of the neural networks.

From the figures presented below, a path with a black line represents a positive path from the node to neuron. A grey line represents a negative path from node to neuron. The width of the line represents the weight of the pathway. For the 50% training model, hidden neurons three and ten present positive pathways to predicting readmission, with the other neurons having negative pathways.

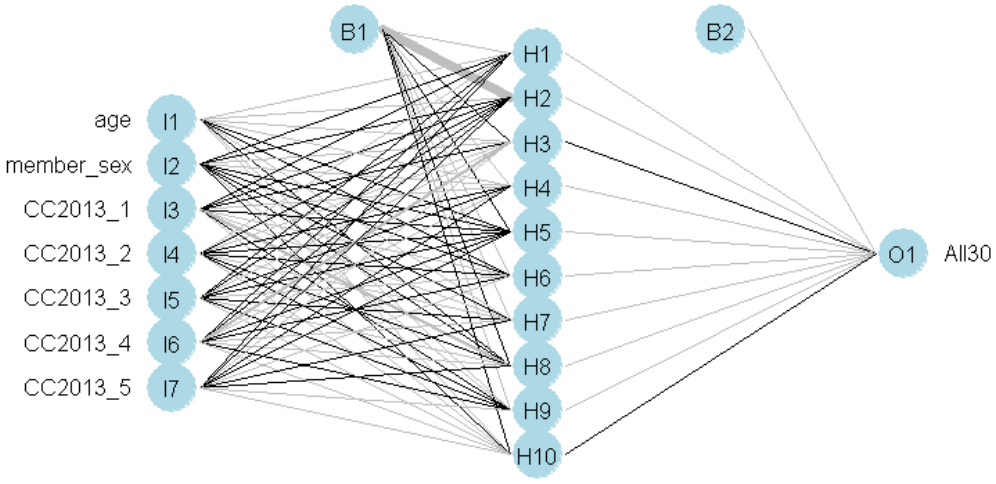


Figure 22. Neural Network Model at 50% Training.

Artificial Neural Networks Graphs for Each Training Level

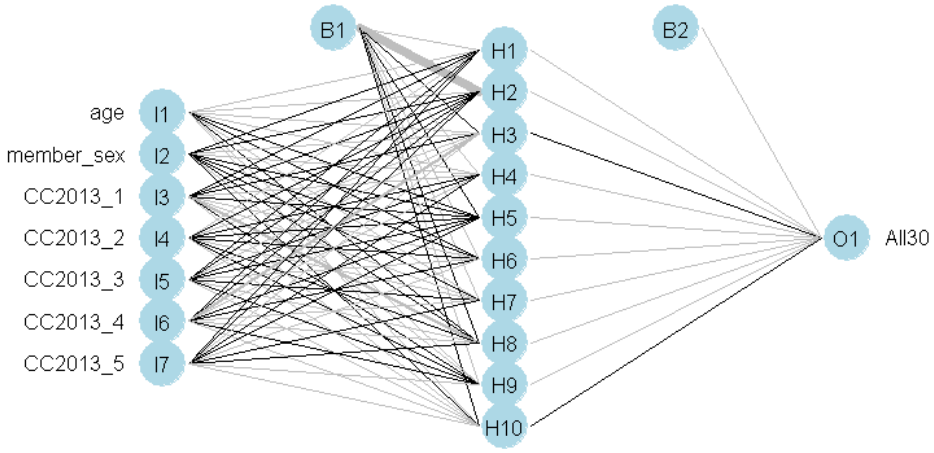


Figure 23. Neural Network Model at 50% Training.

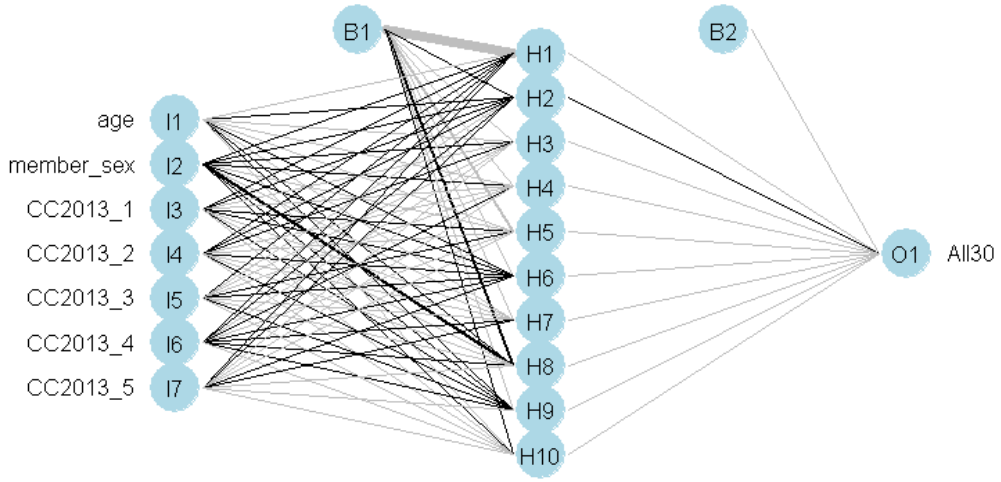


Figure 24. Neural Network Model at 60% Training.

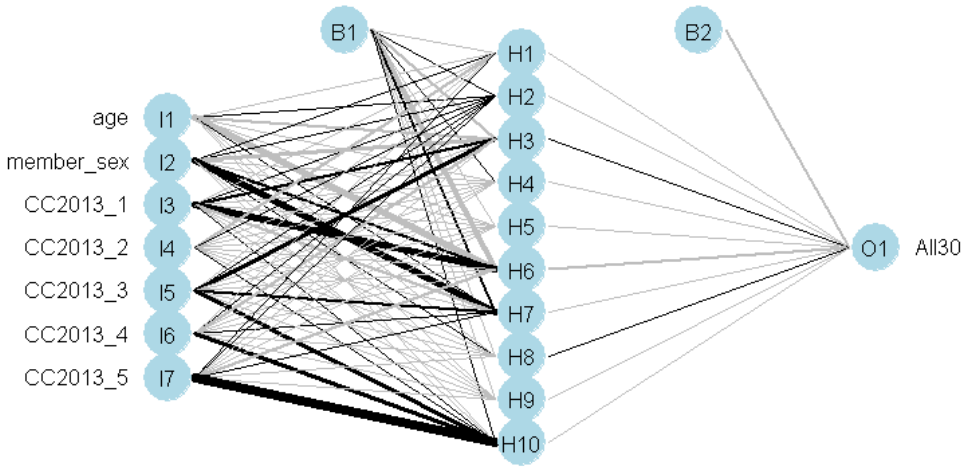


Figure 25. Neural Network Model at 70% Training.

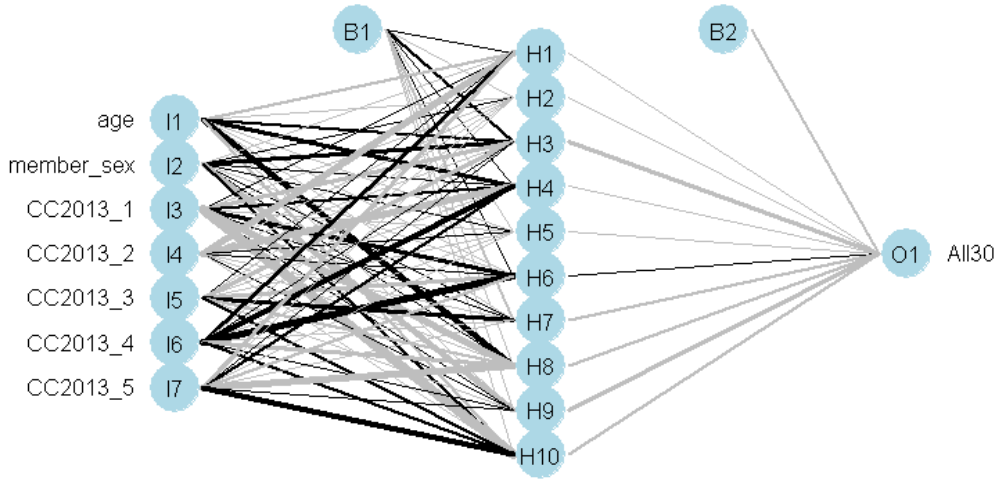


Figure 26. Neural Network Model at 75% Training.

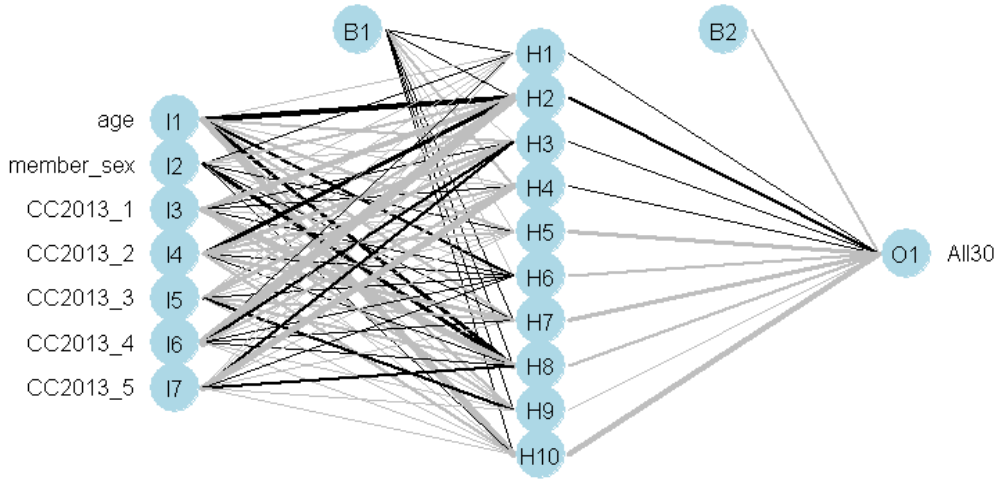


Figure 27. Neural Network Model at 80% Training.

Artificial Neural Network Model Validation

The neural network model was trained and validated on the same dataset as the other analytics. However, the analytic began to decline and overfit at the 60% training interval. As such, it did not perform as well as the decision tree model, but still outperformed the counterpart CMS Logit Model with the same training and validation data.

Table 26. ROC Performance for Neural Networks.

	30Day Readmission	
	Logit Model ROC	Neural Network ROC
50%/50%	0.618	0.798
60%/40%	0.613	0.795
70%/30%	0.617	0.713
75%/25%	0.622	0.780
80%/20%	0.630	0.732

Table 27. Performance Comparison of Neural Networks to CMS Logit Model.

Model	Predictive Accuracy Measures					
	Sensitivity	Specificity	Accuracy	PPV	NPV	AUROC
Baseline	0.521	0.679	0.677	0.024	0.989	0.613
NeuralNet	0.655	0.902	0.899	0.092	0.994	0.798

In the case of the neural networks, the sensitivity of the analytic was far lower than that of the counterpart analytics, but the specificity outperformed the other analytics. This means that the ability of the artificial neural network to take the existing data fields and predict a readmission is more accurate when the admission does not lead to a

readmission. Additionally, the accuracy of the model was contingent on the ability of the network to predict negative values, or non-readmits.

As with the decision tree ROC curve, the neural network shows strong performance in the first two deciles, but with more curve as opposed to immediate lift on sensitivity. In the case of the neural network, this is attributed the ability of the latent nodes in the hidden layer to optimize the outcome based on prediction path. The vast majority of the dataset contains non-readmissions as the outcome, and the neural network performed accordingly.

Artificial Neural Network ROC Curves – All Training Levels

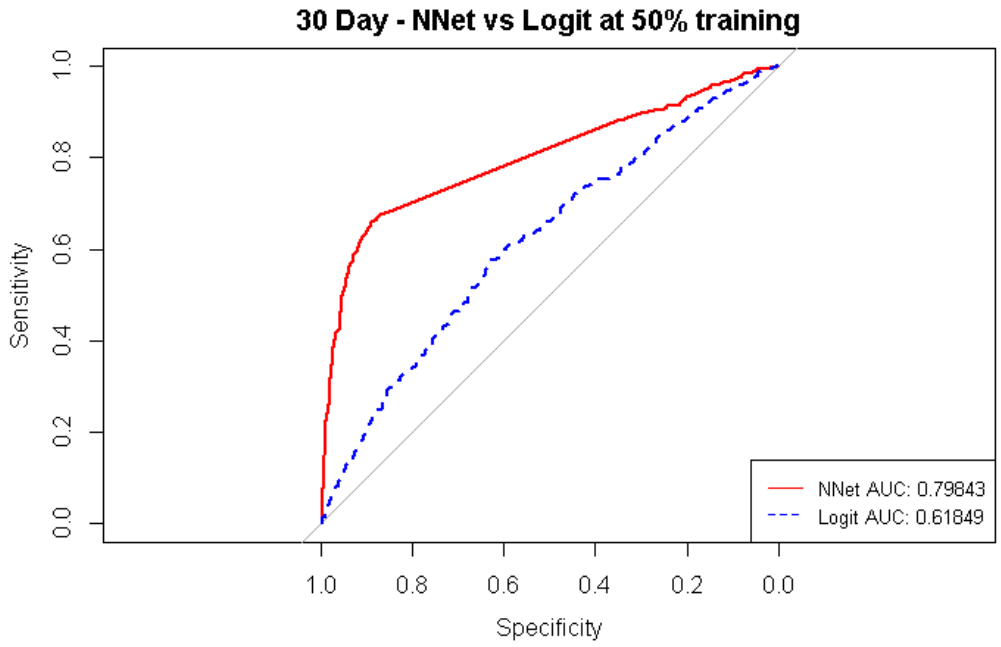


Figure 28. Neural Network ROC for 50% Training.

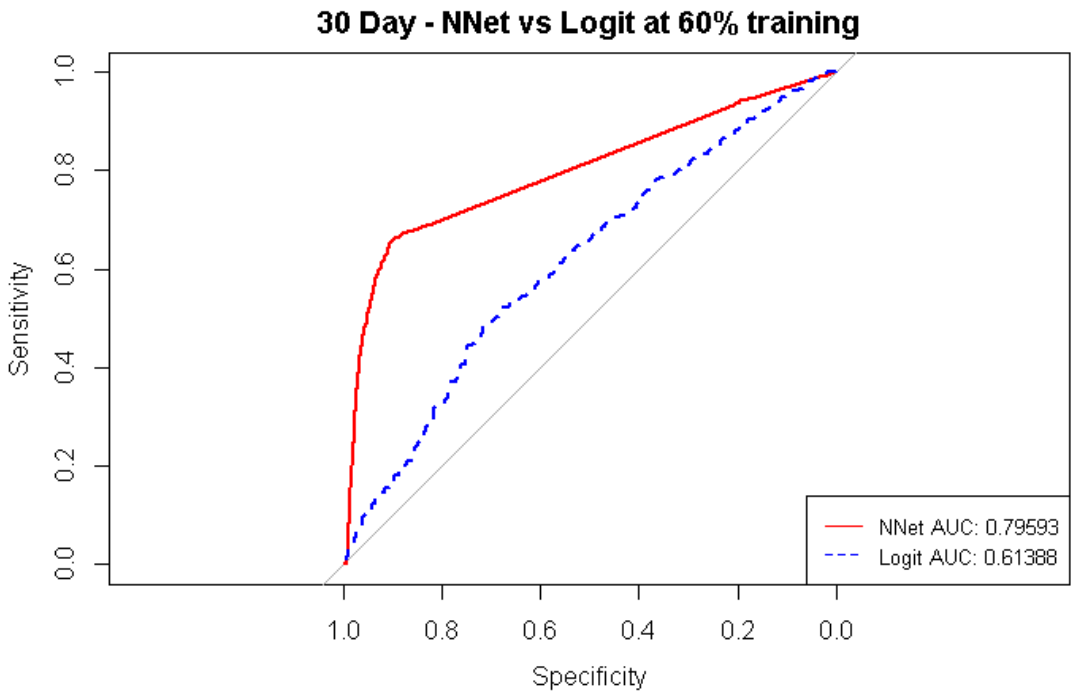


Figure 29. Neural Network ROC for 60% Training.

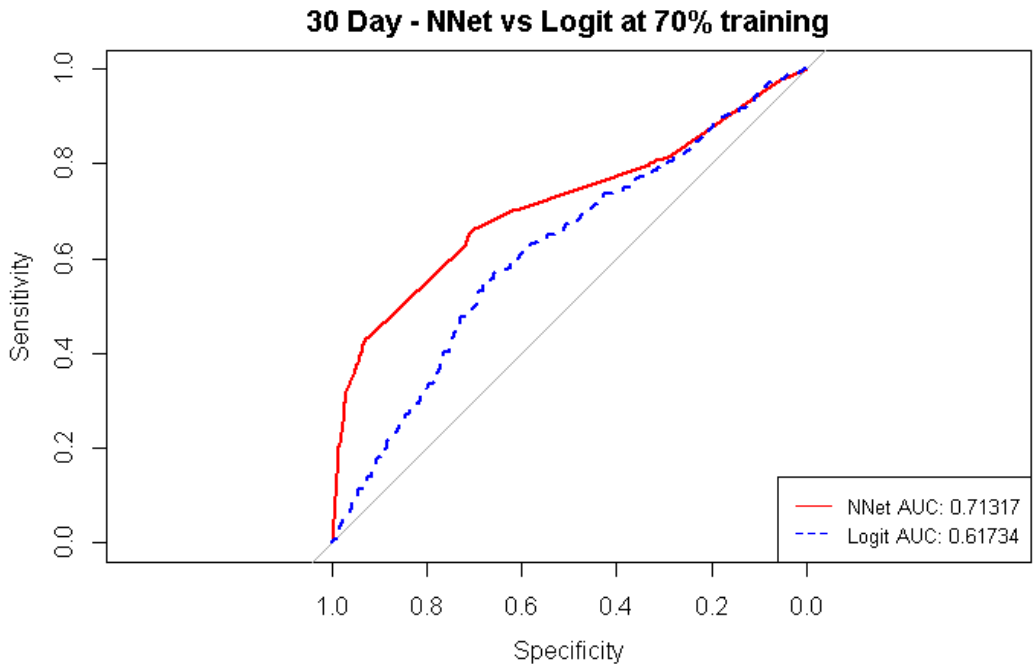


Figure 30. Neural Network ROC for 70% Training.

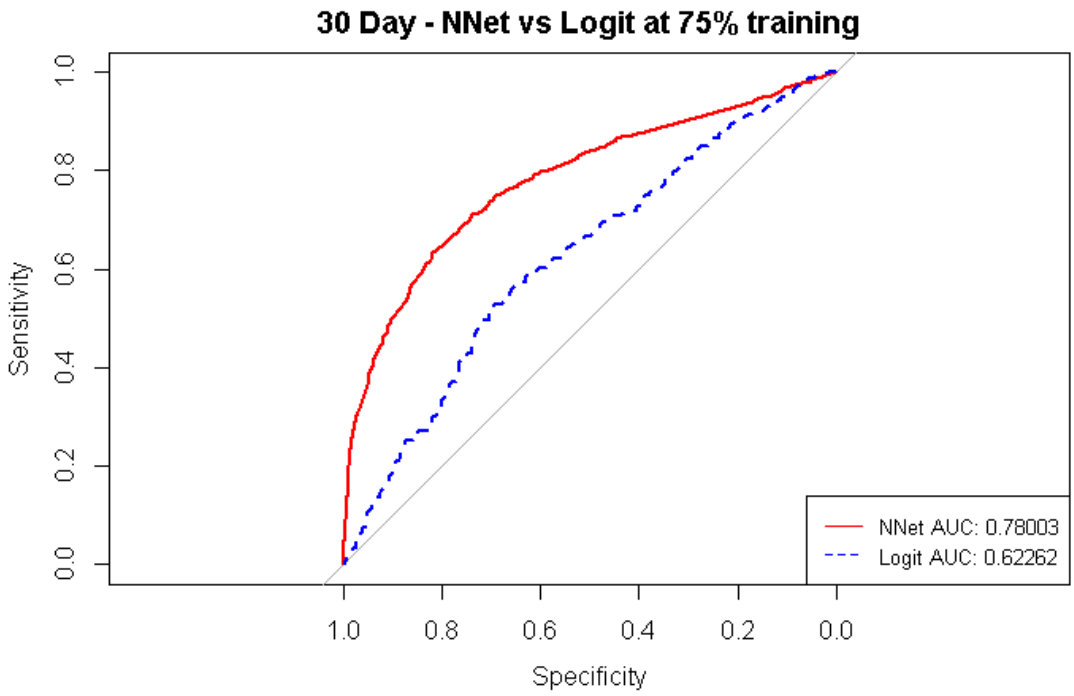


Figure 31. Neural Network ROC for 75% Training.

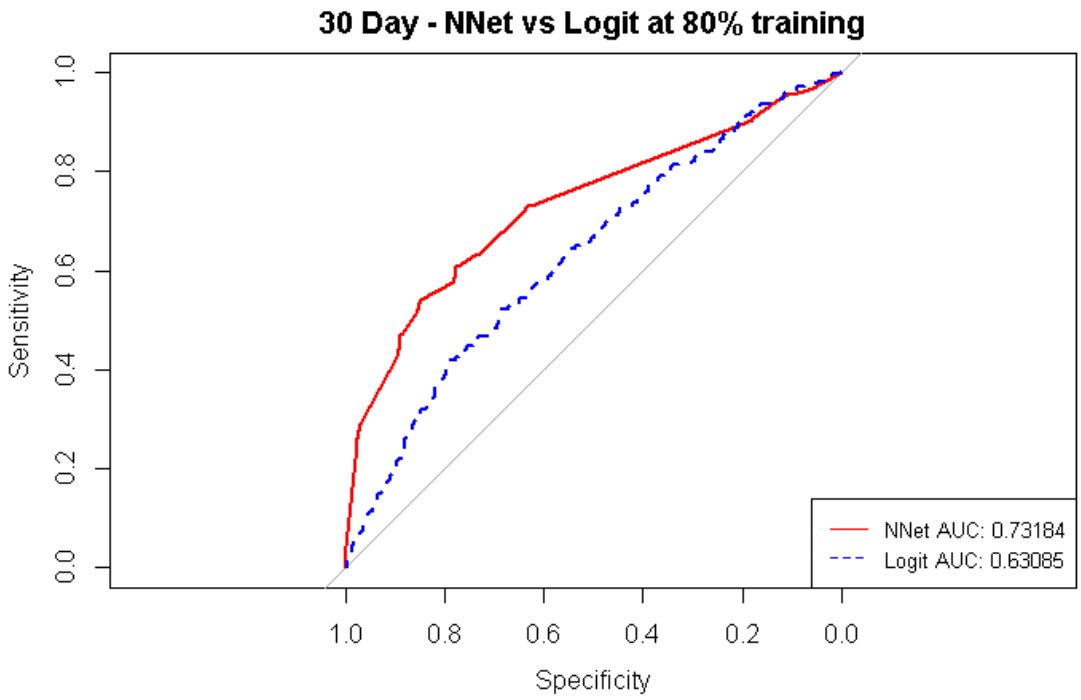


Figure 32. Neural Network ROC for 80% Training.

Self Organizing Maps (SOM) Results

Self organizing maps are a special form of neural network with the intent on reducing dimensionality of data into an x-y grid. This grid is comprised of nodes that examine the relative correlation of data. The algorithm organizes the map based on the adjacency of similarity of attributes. The results color code each column of the data, and the size of the wedge in each node indicates the importance of that attribute to the node. Each node in the map represents a combination of factors and their importance within the node. The position of the node relative to other nodes represents that combination of factors relative to other combination of factors.

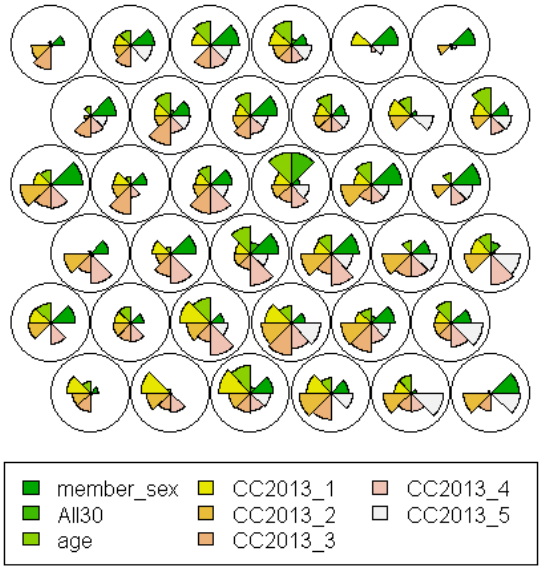


Figure 33. SOM Codes Results - 75% Training

The results of the self organizing map (SOM) algorithm provided the weakest of the results, with only two training levels able to exceed the CMS logit model. The self organizing map results were generated using the R package kohonen (Wehrens and Buydens, 2007). The poor predictive performance SOM is in line with the performance of the neural network. Where the neural network model overfit very rapidly, the SOM algorithm did not begin to find a fit that exceeded the logit model until it was trained on 75% of the data, but then overfit at 80% of the data. However, the algorithm was able to perform better than the logit model with a 75% training/25% validation data set. The results are discussed below.

SOM's purpose is to produce a two dimensional plot of multidimensional data, as referenced in figure 34. Since each circle represents a combination of factors that can predict readmission, the circles that have the highest value for the 'All30,' would be the indicators for readmission. To better understand how related these groupings of characteristics are, SOM utilizes a distance plot. The distance plot can be found in figure 35. The distance plot visualizes the distance of each node from each other when a prediction variable is specified. For this model, the all cause readmission variable was specified. Red in the map indicates nodes that have minimal distance from each other in terms of their prediction based on attributes. There is a grouping of these nodes in the upper middle of the map. These distances suggest that sex, age, readmission, and primary diagnosis are all adjacent to each other.

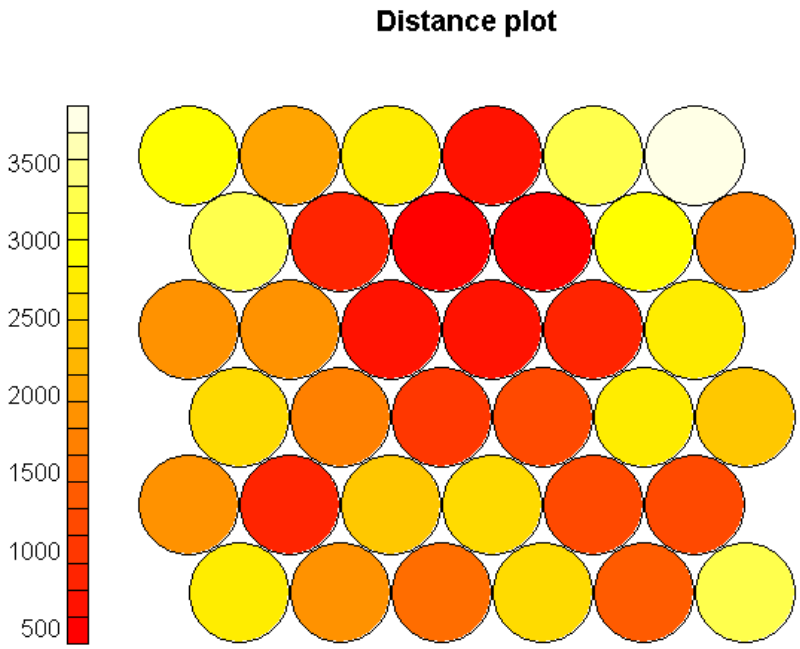


Figure 34. SOM Distance Results - 75% Training

Self Organizing Maps for Each Training Level

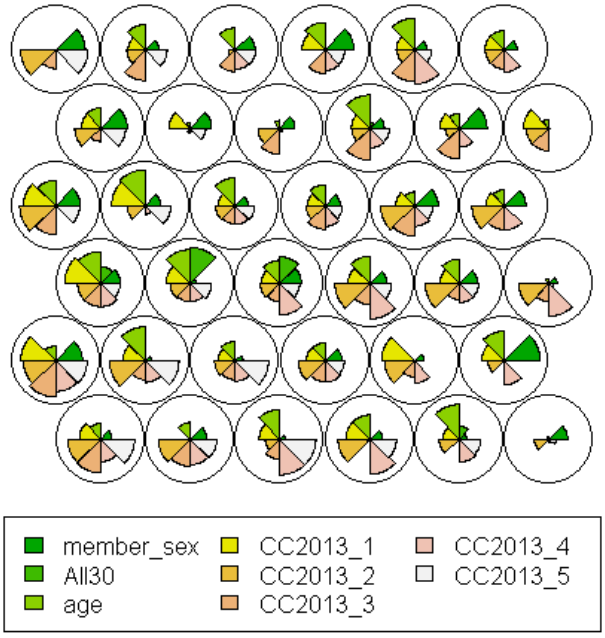


Figure 35. SOM Codes Results - 50% Training

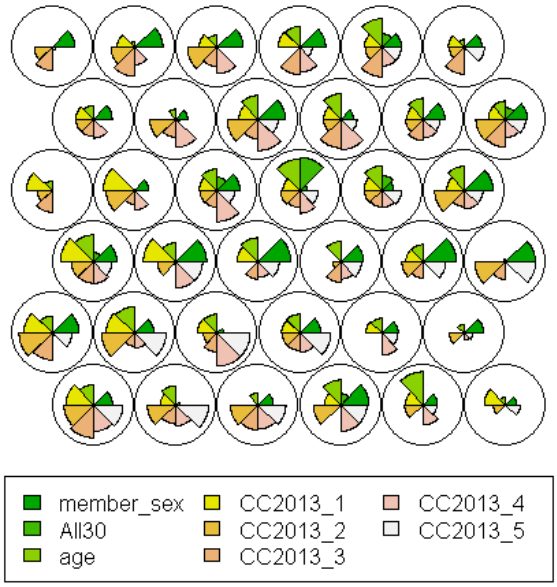


Figure 36. SOM Codes Results - 60% Training

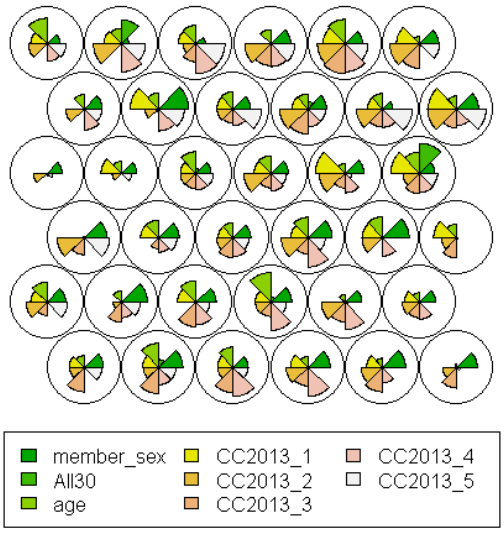


Figure 37. SOM Codes Results - 70% Training

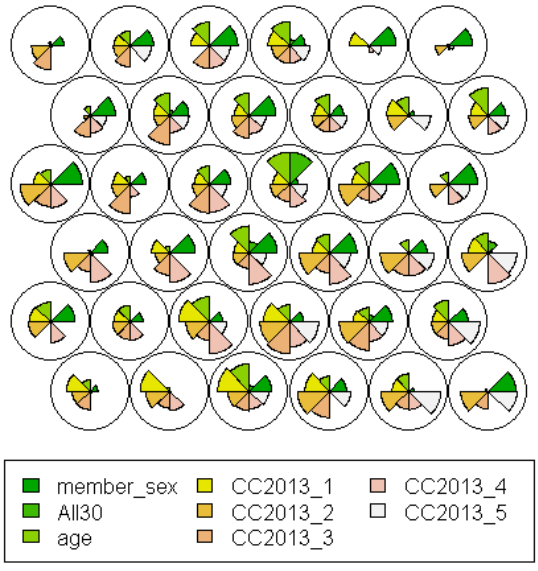


Figure 38. SOM Codes Results - 75% Training

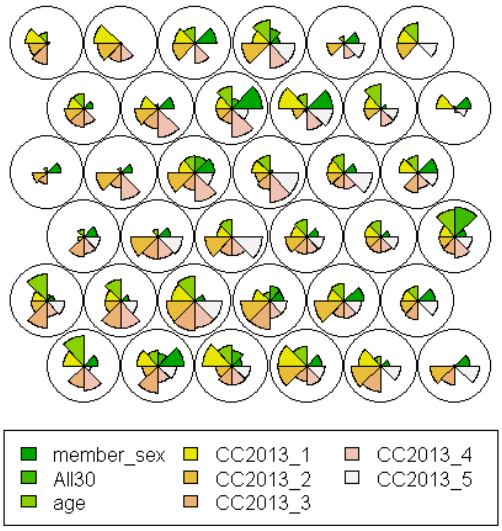


Figure 39. SOM Codes Results - 80% Training

Self Organizing Maps Model Validation

The SOM algorithm's ROC curves are quite different from the other algorithms in that they show plateaus in sensitivity where the others show increases or decreases. This is attributable to the algorithm taking the dimensionality of the data and reducing it to a set of x-y coordinates and predicting a readmission based on the combination of the placement of the node in the map in conjunction with the attributes inside the node. As the algorithm learns from observed positive value, it improves; since the number of observed positive readmits is sparse, the sensitivity plateaus are observed.

Table 28. ROC Performance for SOM.

	30Day Readmission	
	Logit Model ROC	Decision Tree ROC
50%/50%	0.618	0.470
60%/40%	0.613	0.533
70%/30%	0.617	0.524
75%/25%	0.622	0.723
80%/20%	0.630	0.667

Table 29. Performance Comparison of SOM to CMS Logit Model.

Model	Predictive Accuracy Measures					
	Sensitivity	Specificity	Accuracy	PPV	NPV	AUROC
Baseline	0.571	0.650	0.649	0.026	0.989	0.622
SOM	0.667	0.827	0.825	0.047	0.995	0.723

This is further reinforced when examining the sensitivity percentage. This is evidenced by the high specificity to the low sensitivity in the best ROC result, as well as the high negative predicted value. The algorithm was able to outperform the logit model,

but only once before it was over-trained. The SOM algorithm demonstrated the weakest predictive performance of the four analytics being examined.

Self Organizing Maps ROC Curves - All Training Levels

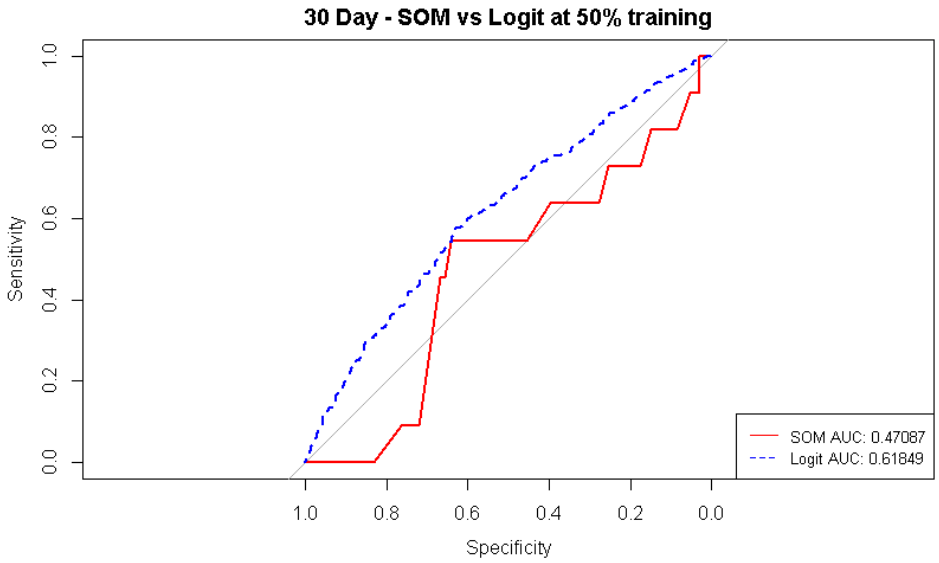


Figure 40. SOM ROC for 50% Training.

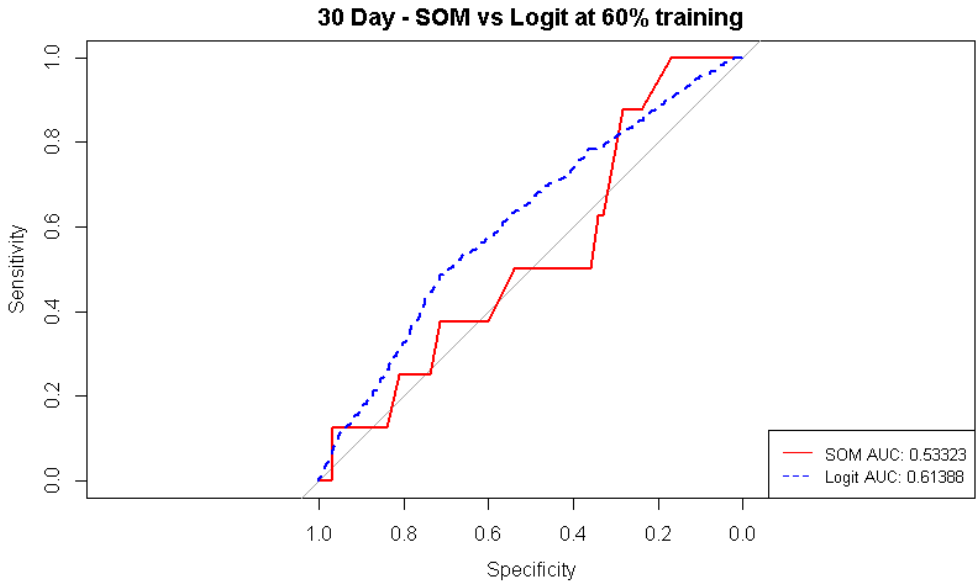


Figure 41. SOM ROC for 60% Training.

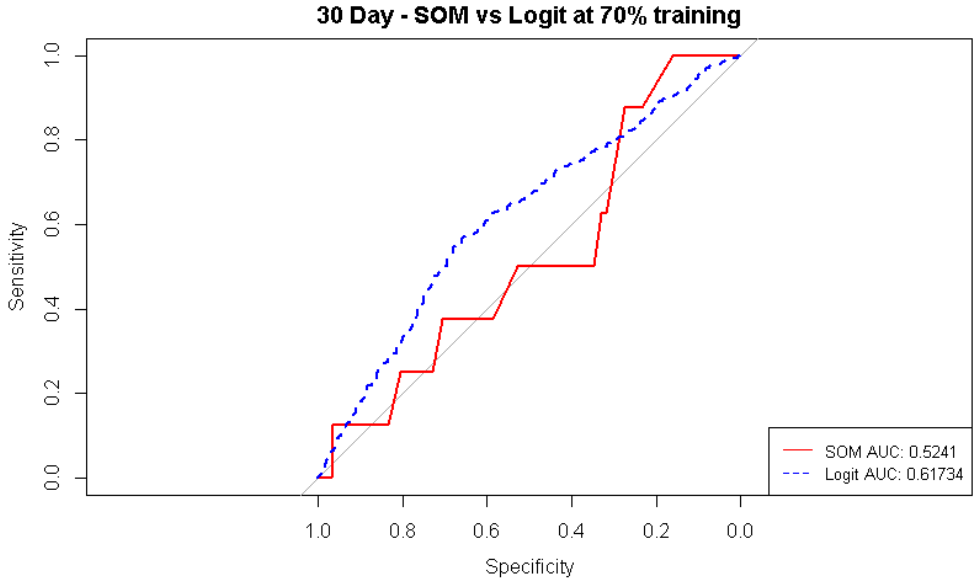


Figure 42. SOM ROC for 70% Training.

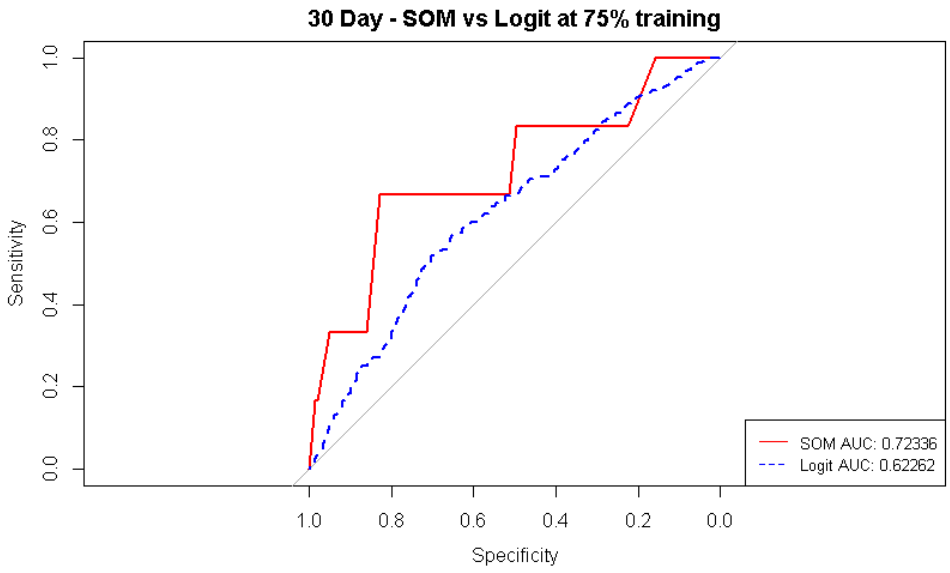


Figure 43. SOM ROC for 75% Training.

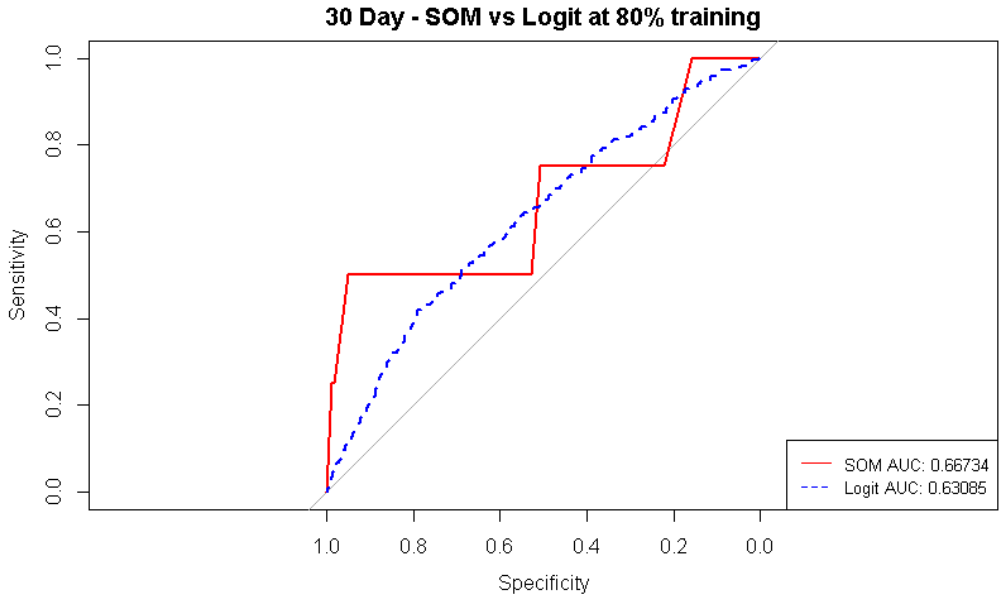


Figure 44. SOM ROC for 80% Training.

Naïve Bayesian Classifier Results

The naïve Bayesian classifier is different in that it is based off of Bayesian probabilities. The classifier examines the probability of each data column's values and then provides probabilities based on the combination of those values. The naïve Bayesian classifier applies Bayes' theorem with a strong independence of assumptions (Naïve) between characteristics in the data. This means that the classifier is resilient to identifying one specific attribute of data and then using it to predict. In the case of this dataset, that attribute would be if a CC is coded with 80 or 131, heart failure and renal failure. The naïve Bayes classifier was modeled using the e1073 package for R (Meyer et al, 2015).

The results that show a probability of a readmission are reported below. These results are interpreted as follows. First, the independent variable is provided. The next three rows represent the dependent variable, with Y, 0 and a 1, with a 0 predicting no readmit and a 1 predicting a readmit.

Table 30. Naive Bayes Classifier Results - 70%.

Age			
Y	[,1]	[,2]	
0	77.178	7.989	
1	83.44444	5.570	
Sex			
Y	0 (Female)	1 (Male)	
0	0.536	0.463	
1	0.888	0.111	
CC2013_1			
Y	78	80	
0	0.065	0.240	
1	0.111	0.555	
CC2013_2			
Y	80	92	131
0	0.295	0.046	0.235
1	0.444	0.333	0.222
CC2013_3			
Y	79	80	131
0	0.069	0.262	0.199
1	0.333	0.222	0.222
CC2013_4			
Y	80	108	131
0	0.319	0.052	0.150
1	0.222	0.444	0.222
CC2013_5			
Y	80	92	108
0	0.303	0.128	0.065
1	0.222	0.222	0.444

The results of the naïve Bayes classifier can be interpreted as follows. The value in the 0 row of any of the categories represents the probability of not being readmitted, while the value in the 1 row of the corresponding category represents the probability of a readmission.

These results suggest that it is far more probable for a female with the corresponding CC codes to be readmitted than for a male. It is not surprising that CC code 80 shows up, as that is the heart failure CC. Additionally CC131 represents renal failure. However, the naïve Bayes classifier suggests that patients with a CC code of 78, respiratory arrested) as their primary diagnosis or 92 (specified heart arrhythmias) as a codiagnosis or comorbidity also present a probable readmission. Lastly, the naïve Bayes classifier provided a result on a comorbidity of CC108, COPD.

Bayesian Probabilities

50% Training

Conditional probabilities:

```
   age
Y       [,1]     [,2]
  0 77.20214 8.035973
  1 81.66667 7.023769

   member_sex
Y          0         1
  0 0.5347594 0.4652406
  1 0.6666667 0.3333333

   CC2013_1
Y           80         130          108
  0 0.249197861 0.013903743 0.0659509202
  1 0.666666667 0.333333333 0.4444444444

   CC2013_2
Y           80
  0 0.291978610
  1 0.333333333

   CC2013_3
Y           80        131
  0 0.274866310 0.207486631
  1 0.333333333 0.666666667

   CC2013_4
Y           80        131
  0 0.330481283 0.145454545
  1 0.333333333 0.333333333

   CC2013_5
Y           80         92
  0 0.298395722 0.127272727
  1 0.333333333 0.666666667
```

60% Training

Conditional probabilities:

```
  age
Y       [,1]     [,2]
  0 77.16637 7.998044
  1 83.00000 6.350853

  member_sex
Y          0         1
  0 0.5339893 0.4660107
  1 0.8571429 0.1428571

  CC2013_1
Y           79          80          92         130         164
  0 0.0644007156 0.2415026834 0.1073345259 0.0152057245 0.0205724508
  1 0.1428571429 0.4285714286 0.1428571429 0.1428571429 0.1428571429


  CC2013_2
Y           80          92         131
  0 0.2915921288 0.0483005367 0.2289803220
  1 0.5714285714 0.1428571429 0.2857142857

  CC2013_3
 Y           9          79          80         131
  0 0.0107334526 0.0706618962 0.2710196780 0.1976744186
  1 0.1428571429 0.1428571429 0.2857142857 0.2857142857

  CC2013_4
Y           19          80         108         131
  0 0.0259391771 0.3220035778 0.0563506261 0.1449016100
  1 0.1428571429 0.2857142857 0.2857142857 0.2857142857

  CC2013_5
Y           80          92         108         131
  0 0.3067978533 0.1279069767 0.0724508050 0.1073345259
  1 0.2857142857 0.2857142857 0.2857142857 0.1428571429
```

75% Training

Conditional probabilities:

```
   age
Y       [,1]     [,2]
  0 77.16774 7.958138
  1 82.58333 5.484828

   member_sex
Y          0         1
  0 0.5333333 0.4666667
  1 0.8333333 0.1666667

   CC2013_1
Y           79          80          92         164
  0 0.0637992832 0.2408602151 0.1096774194 0.0222222222
  1 0.0833333333 0.6666666667 0.0833333333 0.0833333333

   CC2013_2
Y           80          92
  0 0.2924731183 0.0465949821
  1 0.3333333333 0.3333333333

   CC2013_3
Y           80          92         131
  0 0.2630824373 0.0824372760 0.2007168459
  1 0.2500000000 0.0833333333 0.3333333333

   CC2013_4
Y           79          92         108         131
  0 0.0387096774 0.3146953405 0.0530465950 0.1519713262
  1 0.0833333333 0.2500000000 0.3333333333 0.1666666667

   CC2013_5
Y           80          92         108
  0 0.2974910394 0.1318996416 0.0623655914
  1 0.2500000000 0.2500000000 0.3333333333
```

80% Training

Conditional probabilities:

```
   age
Y       [,1]     [,2]
  0 77.26093 8.017509
  1 81.76923 6.016004

   member_sex
Y          0         1
  0 0.5312710 0.4687290
  1 0.7692308 0.2307692

   CC2013_1
Y           79         80         92        108
  0 0.063214526 0.243443174 0.108271688 0.053127102
  1 0.076923077 0.615384615 0.076923077 0.076923077

   CC2013_2
Y           79         80         92        131
  0 0.145931406 0.292535306 0.044384667 0.232010760
  1 0.076923077 0.307692308 0.307692308 0.307692308

   CC2013_3
Y           79         80         92        131
  0 0.072629455 0.268325488 0.084734364 0.195696032
  1 0.230769231 0.230769231 0.076923077 0.384615385

   CC2013_4
Y           19         79         80        108        131         148
  0 0.028917283 0.039677202 0.313382650 0.051782112 0.154673840 0.014794889
  1 0.076923077 0.076923077 0.307692308 0.307692308 0.153846154 0.076923077

   CC2013_5
Y           80         92        108         131
  0 0.289845326 0.131809011 0.065232011 0.108271688
  1 0.307692308 0.230769231 0.307692308 0.153846154
```

Bayesian Classifier Model Validation

The naïve Bayesian classifier performed second best, next to the decision trees. Overfitting began to occur beyond the 70% training interval. The ROC curves for the Bayes classifier are striking compared to the other results when examining the sensitivity in the first decile. The Bayes classifier demonstrates a straight lift, which is attributable to the probabilistic nature of analytic. This shows a different view of lift than the SOM map, in that the naïve Bayes classifier is looking for predicted positive values, and it is shown here that the PPV of the Bayes classifier is so much higher than the baseline.

Table 31. ROC Performance for Naive Bayes Classifier.

	30 Day Readmission	
	Logit Model ROC	Bayesian ROC
50%/50%	0.618	0.666
60%/40%	0.613	0.713
70%/30%	0.617	0.721
75%/25%	0.622	0.623
80%/20%	0.630	0.618

Table 32. Performance Comparison of Naive Bayes Classifier to CMS Logit Model.

Model	Predictive Accuracy Measures					
	Sensitivity	Specificity	Accuracy	PPV	NPV	AUROC
Baseline	0.571	0.658	0.656	0.021	0.989	0.617
Bayes	0.444	0.998	0.993	0.571	0.996	0.721

As with the other analytics, the naïve Bayes classifier performed poorly with a sparse number of readmits. However, it performed the best with the number of non-readmits.. While the sensitivity of the classifier underperformed compared to the baseline, the specificity and accuracy were very high.

Bayesian Classifier ROC Curves – All Training Levels

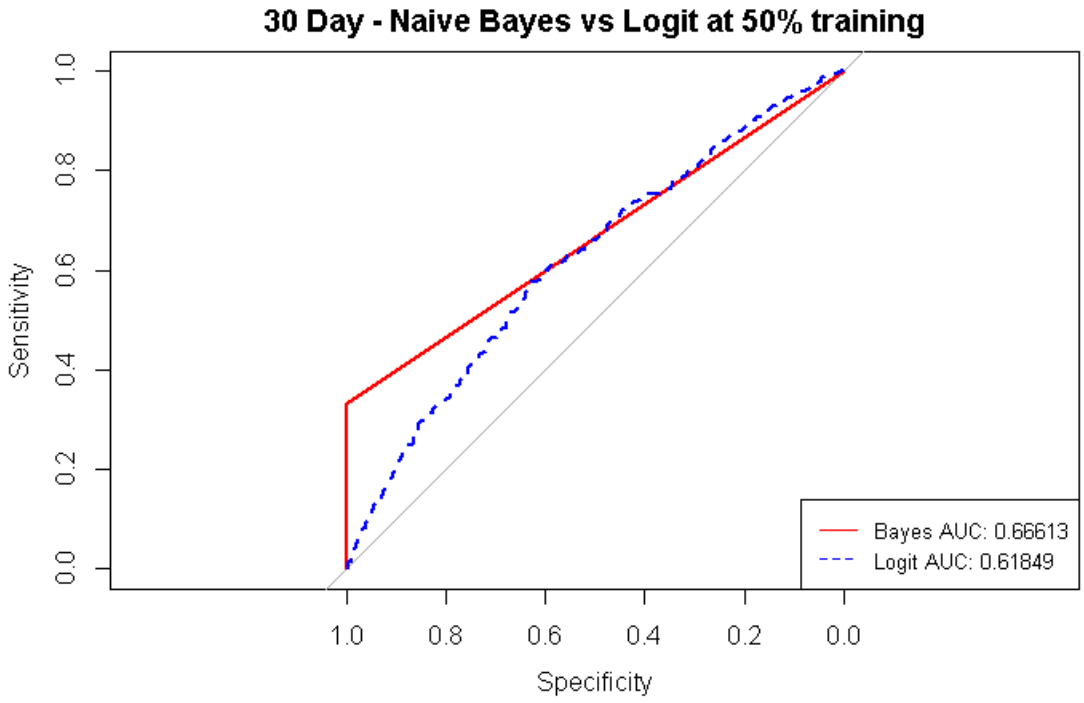


Figure 45. Naive Bayes ROC for 50% Training.

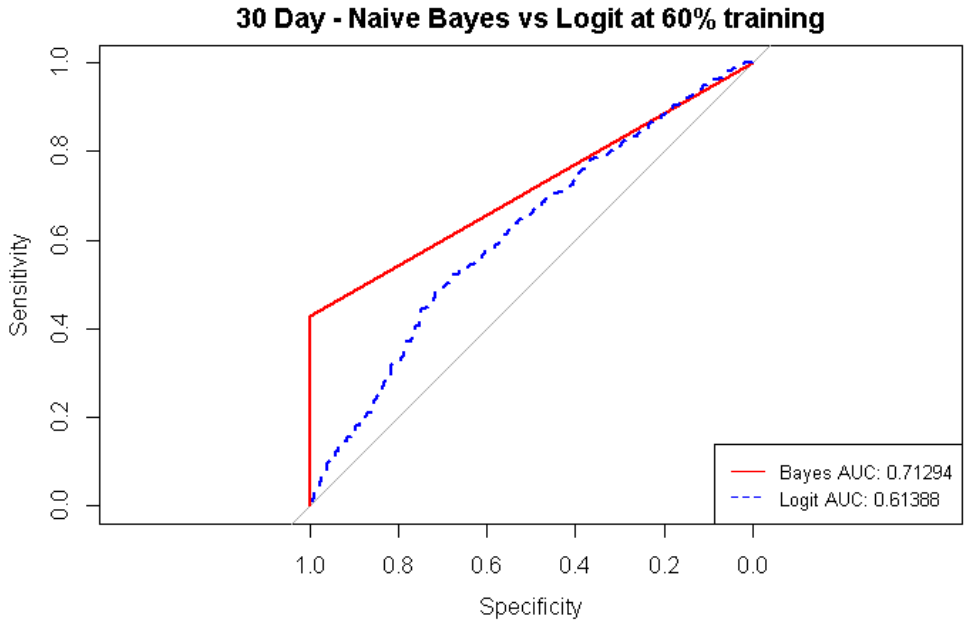


Figure 46. Naive Bayes ROC for 60% Training.

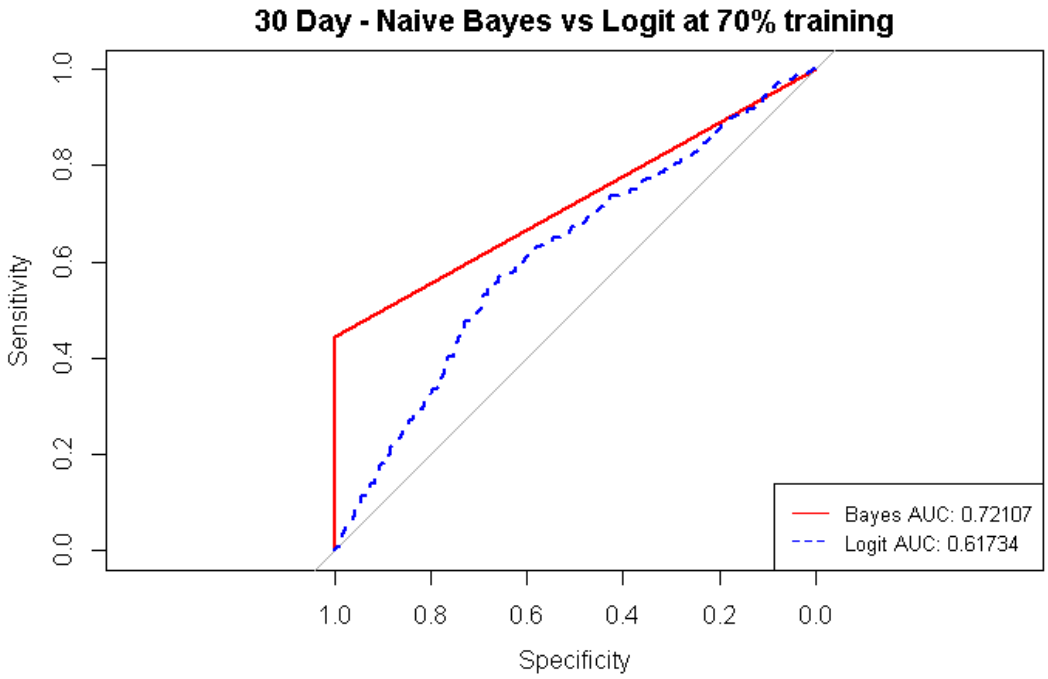


Figure 47. Naive Bayes ROC for 70% Training.

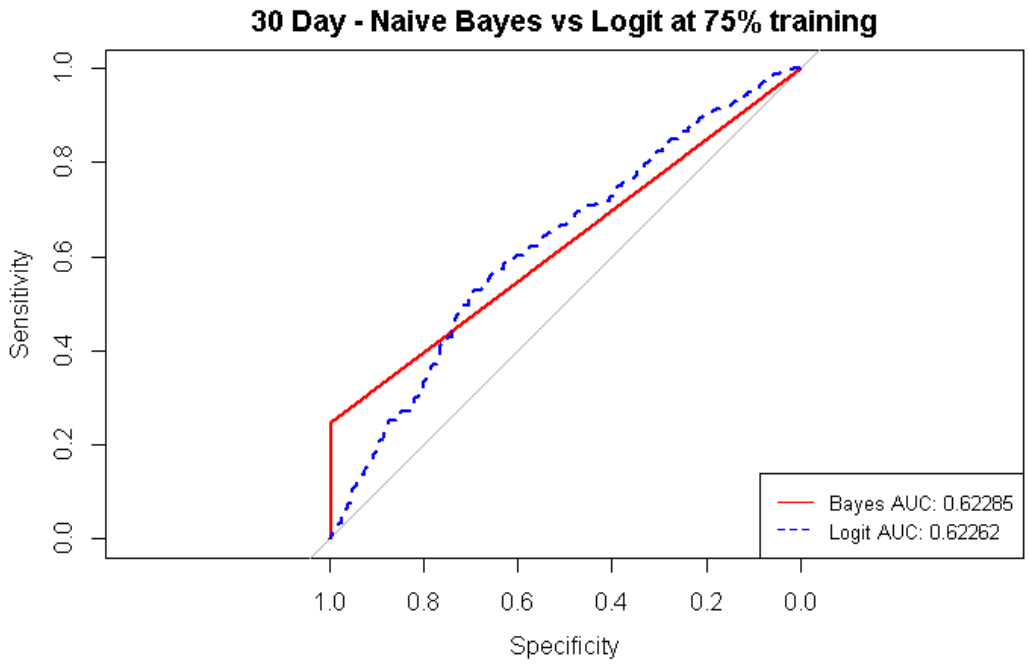


Figure 48. Naive Bayes ROC for 75% Training.

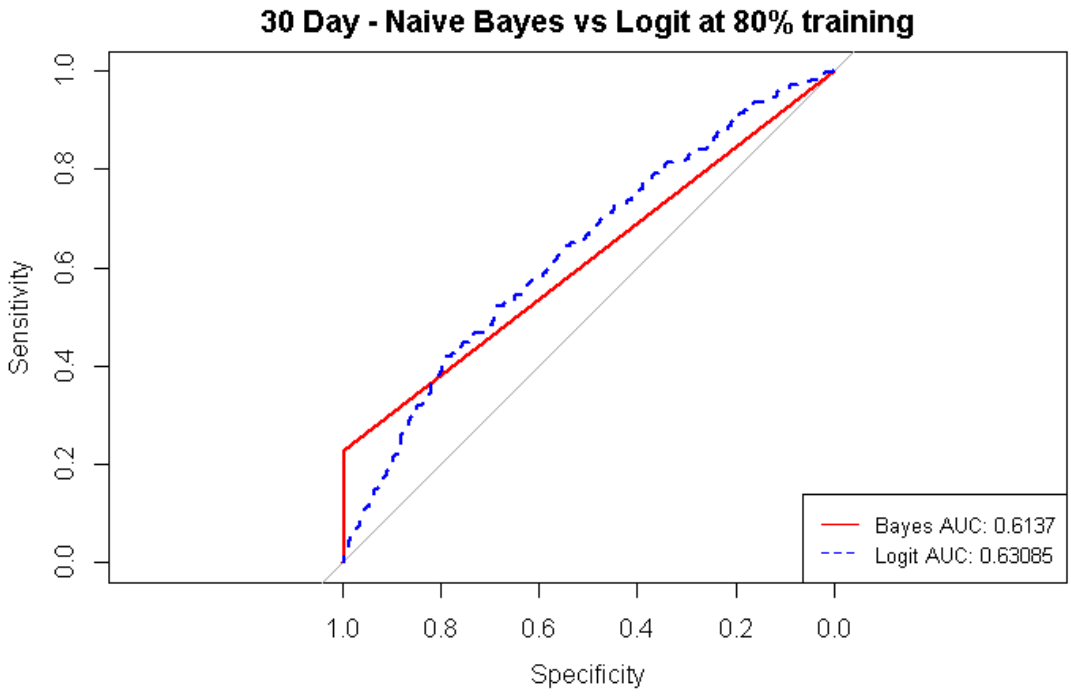


Figure 49. Naive Bayes ROC for 80% Training.

Summary of Analytics Results

This chapter presents the output of four advanced analytics and compares their predictive power at five different training/validation intervals. A summary of the best prediction power of each analytic is found in table 32. Overall, the classification decision tree performed the best, followed by the neural network analytic. Each analytic provides a different view of the data, as well as different prediction abilities. Every analytic was able to outperform the logit model at various training levels, even though some fell behind at a higher training level.

The decision tree analytic expanded the current understanding of readmissions within this dataset by providing a set of decision rules to reach a node in which the majority of the population at the terminal node was readmitted. The neural network was able to provide a better regression model prediction than the logit model while also showing the influence of the data categories on the prediction ability, as well as the bias of the categories on the outcome. The self organizing map is able to plot the adjacency of every category of data as to how they relate to each other, and provide a visual of the influence of each data category to each other data category. The naïve Bayes classifier showed that a probabilistic model, while underperforming compared to deterministic models, was still able to provide insight into the specifics of the CC codes that contribute to predicting readmissions.

Table 33. Summary of Best ROC per Analytic.

Analytic	**Best ROC**	**Training Level**
GLM	0.630	80%
Classification Decision Tree	**0.819**	**75%**
Neural Network	0.798	50%
Self Organizing Map	0.723	75%
Naïve Bayes	0.721	70%

Table 34. Summary of ROC Performance.

	Logit Model	Decision Tree	Neural Net	SOM	Naïve Bayes
50%/50%	0.618	0.795	**0.798**	0.470	0.666
60%/40%	0.613	0.793	**0.795**	0.533	0.713
70%/30%	0.617	**0.817**	0.713	0.524	0.721
75%/25%	0.622	**0.819**	0.780	0.723	0.623
80%/20%	0.630	**0.808**	0.732	0.667	0.618

Whereas the goal of this chapter was to demonstrate improvement of the prediction power of the CMS model, the next chapter's goal is to demonstrate expansion of understanding of the factors that contribute to variances in heart failure readmissions. The next chapter examines a special case of analytics. Cluster modeling has been performed with the data in order to further improve the accuracy of prediction of readmissions.

CHAPTER VI

CLUSTERING RESULTS

This chapter presents the results of the analytics after applying two stage clustering to the NCSHP data. Each of the four analytics from Chapter V is replicated in this chapter. The structure of this chapter is as follows. First, the results of clustering are examined. The results of the cluster membership are described. Next, each analytic from Chapter V are examined after applying cluster membership. The performance of each model is then assessed using the ROC curve. The chapter concludes with a summary of the best performing analytic when incorporating cluster membership.

Clustering Results

The advantage of applying clustering to the data is that it pre-classifies the data into segments. Each cluster represents similar observations. By applying clustering using the categorical condition codes with the all cause readmission flag, clusters form around the condition codes that are more associated with readmissions.

Two step clustering was chosen over hierarchical clustering and kmeans clustering. K-means clustering is not applicable when dealing with categorical data; hierarchical clustering is the second component in the two-step clustering method. Two-step clustering iterates through a pre-defined number of clusters and then verifies the membership through the log-likihood distance of hierarchical membership. For this

dataset, a maximum cluster was set to thirty. The results of the two step clustering yielded eight clusters. The results of these memberships are visualized in figure 53. For the clustering membership, age, sex and CC80 variables were dropped. Age and sex were dropped given clustering's tendency to agglomerate on these variables. Since this dissertation is examining the conditional category profile of patients at risk of a heart failure readmission, the information gain from age and sex would dilute the cluster membership. CC80 is the condition category for heart failure, which all observations contained. The resulting clusters are solely based on condition categories as predictors of heart failure readmission.

The cluster membership was validated using the silhouette method of measuring cohesion and separation. This method is the default evaluation method for the two step clustering algorithm. The silhouette method evaluates how similar an observation is to its own cluster, referred to as cohesion, against how different the observation is compared to other clusters, referred to as its separation. The average silhouette, as referenced in Figure 2, was 0.5. This falls within the -1.0 to 1.0 range and represents decent clustering.

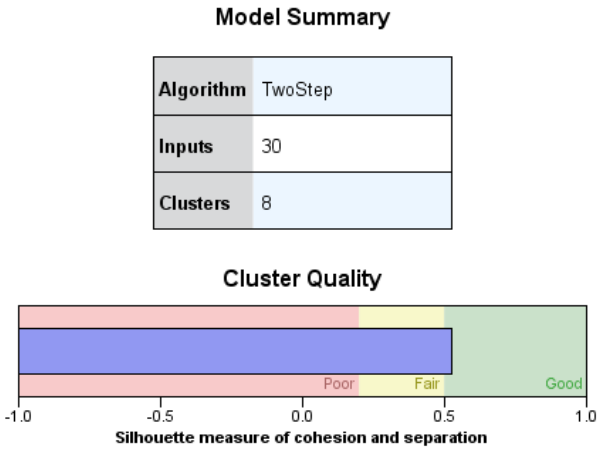
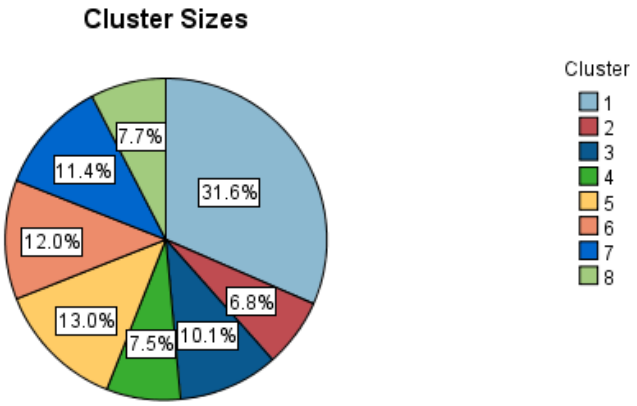


Figure 50. Silhouette Measure of Cluster Separation.



Size of Smallest Cluster	6822 (6.8%)
Size of Largest Cluster	31830 (31.6%)
Ratio of Sizes: Largest Cluster to Smallest Cluster	4.67

Figure 51. Cluster Sizes.

The CC code results of the clustering are presented in figure 53. Cluster one is a cluster in which CC codes were present, and as such, the algorithm grouped those observations together. Cluster two has a very clearly defined grouping around CC108, Chronic Obstructive Pulmonary Disease. Cluster three has a very clearly defined grouping around CC92, Specified Heart Arrhythmias. Cluster four has three clear conditions within its cluster, with CC79: Cardio-respiratory Failure and Shock making up the majority, but also containing CC108, as well as CC131: renal failure. Cluster five is the second largest cluster after cluster one, containing a variety of condition categories,

but very few readmissions. Cluster six has clustered around CC131 and CC92. Cluster seven has clustered around the All30 readmission variable. Cluster eight has clustered around CC19, diabetes without complication.

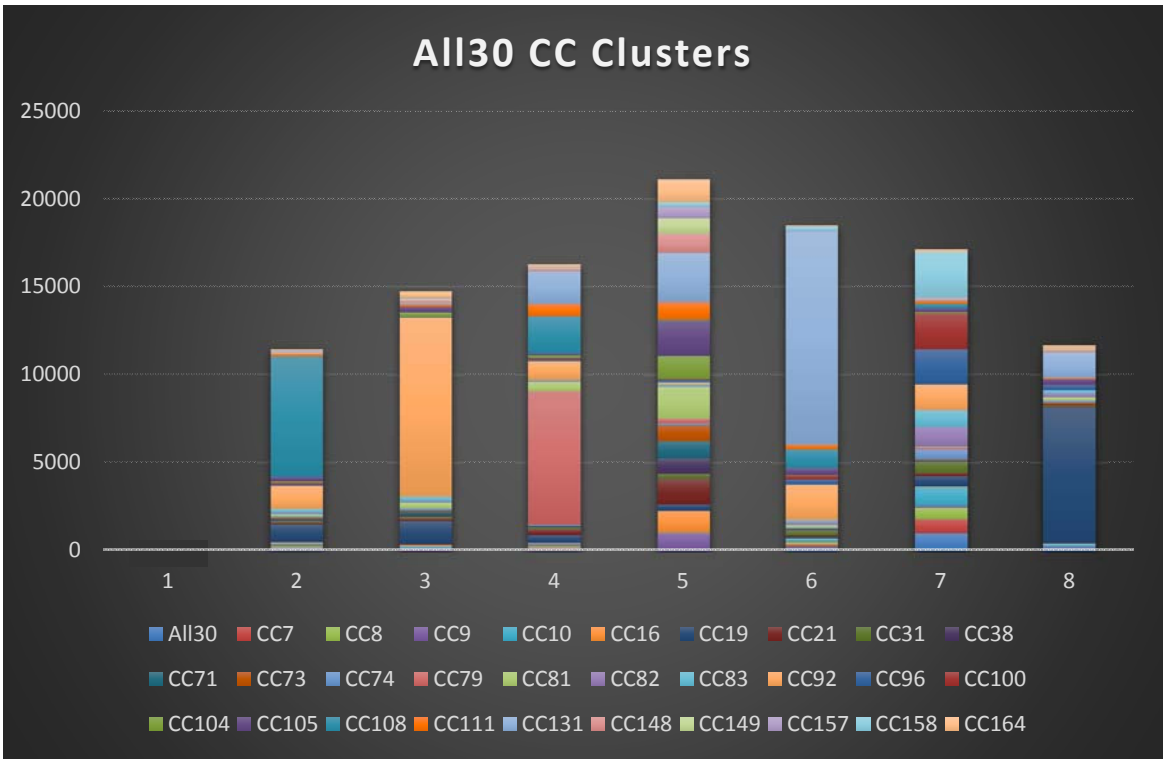


Figure 52. Cluster Membership Based on CC Codes.

From these cluster memberships, cluster seven represents a clear cluster around the All30 variable, which indicates a readmission. This cluster contains the CC codes found in table 35.

Table 35. CC Codes for Cluster 7.

Conditional Category	Description
CC7	Metastatic Cancer and Acute Leukemia
CC8	Lung, Upper Digestive Tract, and Other Severe Cancers
CC10	Breast, Prostate, Colorectal, and other Cancers and Tumors
CC19	Diabetes without Complication
CC31	Intestinal Obstruction/Perforation
CC74	Seizure Disorders and Convulsions
CC81	Acute Myocardial Infarction
CC82	Unstable Angina and Other Acute Ischemic Heart Disease
CC83	Angina Pectoris/Old Myocardial Infraction
CC92	Specified Heart Arrhythmias
CC96	Ischemic or Unspecified Stroke
CC100	Hemiplegia/Hemiparesis
CC105	Vascular Disease
CC108	Chronic Obstructive Pulmonary Disease
CC158	Hip Fracture/Dislocation

The All30 variable in cluster seven contains over half of the total readmission observations. Effectively, this cluster represents a cluster of patients who have been readmitted and presents new information around the condition category profile of those patients. By applying clustering to the data and separating the data based on cluster membership, the data can be modeled again using the analytics from Chapter V. Since cluster seven is the cluster which contains the readmission variable, the analytics have been reassessed using cluster seven's subset of the data to model readmission. The original model predicted readmission as a function of age, sex, and diagnoses codes in the form of CC codes. The clustering model predicts readmission in the form of CC code membership. The follow sections present the results of the analytics using cluster seven.

Decision Tree Results

The classification and decision tree algorithm using the party package in the R platform was used to model cluster seven. The five training and validation levels were again employed to test each algorithm to assess performance before overfitting. These levels were 50%, 60%, 70%, 75%, and 80%. The algorithm performed best at the 80% training/20% validation level.

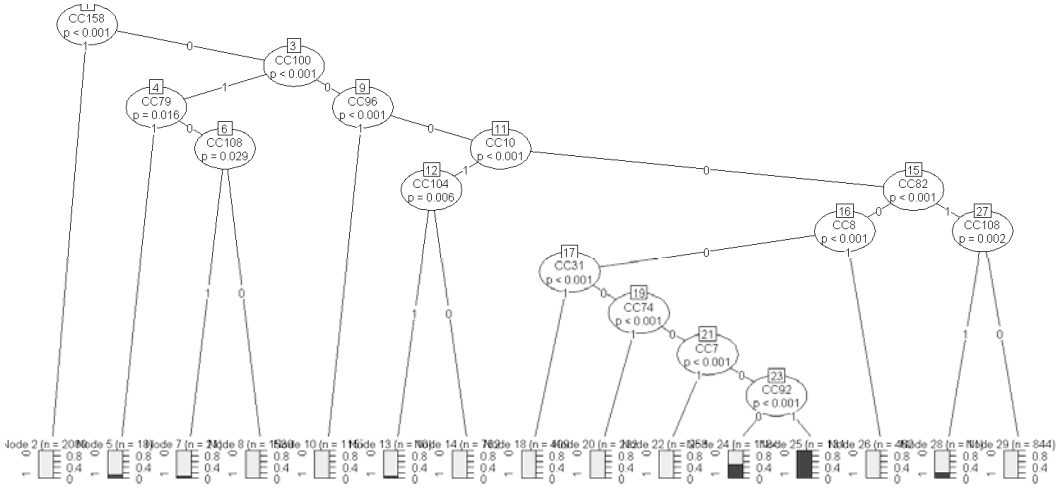


Figure 53. Decision Tree + Clustering at 80% Training.

While reading the decision tree follows the same logic from Chapter V, the presentation of these results is slightly different. The decision tree uses the 0 and 1 paths to assess if a patient has a specific condition. The patient is coded with a 0 if they do not have the condition, and 1 if they do. Node 23 on the decision tree represents the node that contains the most readmits. To reach this node, the follow decisions were applied:

1. The patient does not have a hip/fracture
2. The patient does not have hemiplegia/hemiparesis
3. The patient does not have Ischemic or unspecified stroke
4. The patient does not have breast, prostate, colorectal, and other cancers
5. The patient does not have unstable angina
6. The patient does not have intestinal obstruction/perforation
7. The patient does not have Seizure disorders
8. The patient does not have metastatic cancer and acute leukemia

9. The patient does have specified heart arrhythmias

This tree effectively shows that the most important criteria for predicting heart failure readmissions within this data set are heart arrhythmias. This holds to medical knowledge as well, but with new knowledge around a set of decisions that lead to predicting readmission. The next section presents the decision trees from each training level.

Decision Tree Validation

The result of partitioning the data using clustering led to a stable increase in the ROC area under the curve at each training level. As such, the model did not overfit within the specified training levels for this dissertation when applying clustering. While the decision tree without clustering overfit after 75%, using clustering enabled more accurate predictions with more data.

Table 36. ROC Performance for Decision Trees with Clustering.

	30Day Readmission	
Train%/Test%	Decision Tree ROC	Decision Tree + Clustering ROC
50%/50%	0.795	0.907
60%/40%	0.793	0.907
70%/30%	0.817	0.909
75%/25%	0.819	0.909
80%/20%	**0.808**	**0.917**

Table 37. Performance Comparison of Decision Trees with Clustering.

Model	**Predictive Accuracy Measures**					
	Sensitivity	Specificity	Accuracy	PPV	NPV	AUROC
Dtree	0.720	0.864	0.862	0.077	0.995	0.819
Dtree + Clustering	0.899	0.935	0.932	0.554	0.990	0.917

Both the sensitivity and specificity outperform the original decision tree model. While the sensitivity can be attributable to a higher positive predicted value, the negative predicted value was lower with the cluster membership, but still had a higher specificity than the original decision tree.

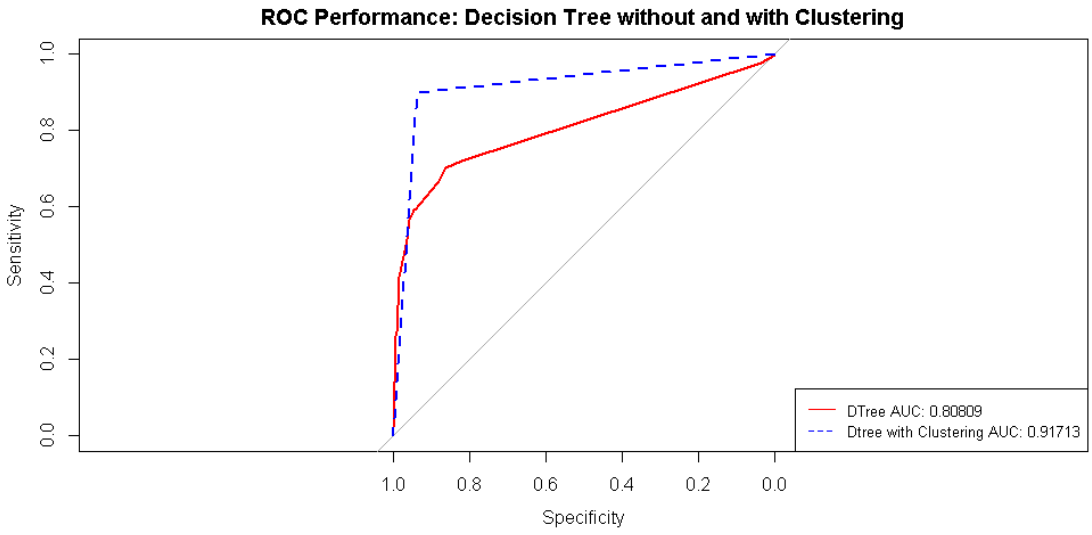


Figure 54. ROC Performance of Decision Trees.

Neural Net Results

The neural network's performance increased substantially when modeling readmission as a function of CC codes for cluster seven. The model was optimized with the 70% training and 30% validation.

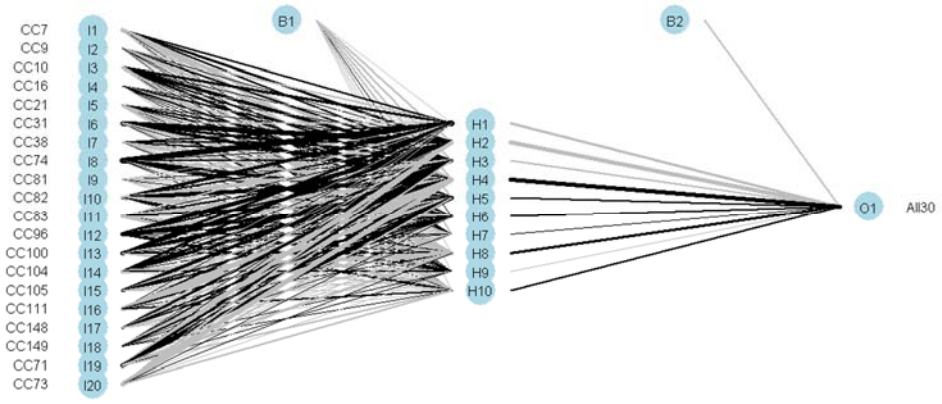


Figure 55. Neural Network + Clustering Results at 70% Training.

The model was specified using the same number of hidden layers and neurons as the original model. The initial nodes were modeled using the CC codes; it is visible that CC96 and CC74 have larger pathways than the other CCs. It is also of note that there are many more positive pathways than the original neural network. Each pathway also has more weight. From the hidden layer, there are also more positive pathways, meaning that the model was predicting more readmits than non-readmits.

Neural Net Validation

The neural network algorithm was able to increase the predictive ability without overfit up to the 70% training interval. While the original neural network provided somewhat unstable ROC results, the clustered model using neural networks performed relatively stable throughout each training level.

Table 38. ROC Performance for Neural Networks with Clustering.

	30Day Readmission	
Train%/Test%	Neural Network ROC	Neural Network + Clustering ROC
50%/50%	0.798	0.818
60%/40%	0.795	0.818
70%/30%	**0.713**	**0.827**
75%/25%	0.780	0.823
80%/20%	0.732	0.818

The original neural network had a far higher specificity than the clustered neural network meaning that it could predict non-readmits with more accuracy. However, the clustered neural network performed at a far higher sensitivity rating, meaning it was able to predict readmissions more accurately. The accuracy of the clustered model is lower, but this is attributable to the poorer performance of the neural network in terms of identifying non-readmits. This is evidenced from the ROC plot, showing a steeper sensitivity rate as specificity levels out.

Table 39. Performance Comparison of Neural Networks with Clustering.

Model	**Predictive Accuracy Measures**					
	Sensitivity	Specificity	Accuracy	PPV	NPV	AUROC
NNet	0.655	0.902	0.899	0.092	0.994	0.798
NNet + clustering	0.881	0.747	0.758	0.245	0.985	0.827

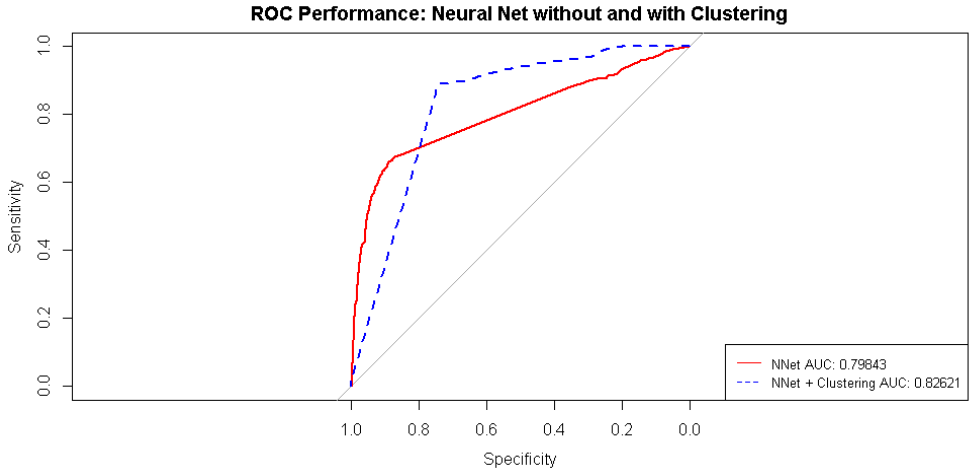


Figure 56. ROC Performance of Clustering + Neural Networks.

SOM Results

The self organizing map algorithm performed best at the 75%/25% training/validation level. The plot shows the reduced dimensions of the data but assessing 11 of the CC codes found in cluster seven. As with the original result, this mapping creates nodes of similar data. Each node represents correlated data, and the position of each node represents how similar these correlations are to each other. The All30 variable is closest related to the CC92 code, which reinforces the findings of heart arrhythmias. Adjacent to this node is the CC8 code (lung, upper digestive tract, and other severe cancers), as well as the CC7 code (metastatic cancer and acute cancer). This map provides additional information beyond the readmission variable, in that each node also groups together various CC codes as to how they relate to each other. For example, CC100 and CC96 demonstrate similarities across the bottom of the map.

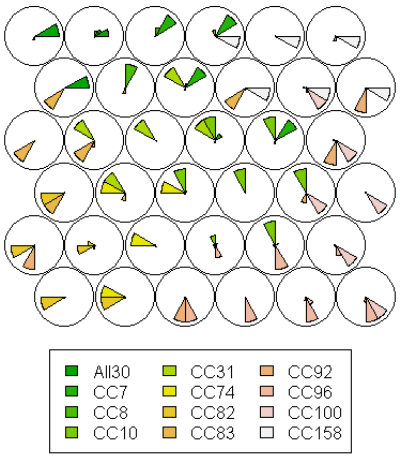


Figure 57. Clustering + SOM Plot for 75% Training

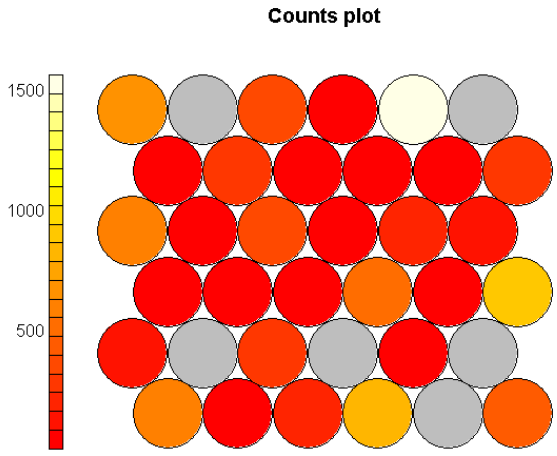


Figure 58. Distance Plot of Clustering + SOM Results.

SOM Validation

The SOM algorithm showed its best performance at the 75% training level. The results from this analytic show a 1.000 sensitivity, meaning the map was able to predict 100% of the readmissions. The map falls short on specificity, in that it had multiple false positives as well for readmission, and was not able to properly separate out non-readmits. However, the sensitivity of the analytic was able to lift the ROC above the prior performance of SOM without clustering.

Table 40. ROC Performance for SOM with Clustering.

	30Day Readmission	
Train%/Test%	SOM ROC	SOM + Clustering ROC
50%/50%	0.470	0.662
60%/40%	0.533	0.706
70%/30%	0.524	0.859
75%/25%	0.723	0.892
80%/20%	0.667	0.850

Table 41. Performance Comparison of SOM with Clustering.

Model	Predictive Accuracy Measures					
	Sensitivity	Specificity	Accuracy	PPV	NPV	AUROC
SOM	0.667	0.827	0.825	0.047	0.995	0.723
SOM + Clustering	1.000	0.783	0.802	0.305	1.000	0.892

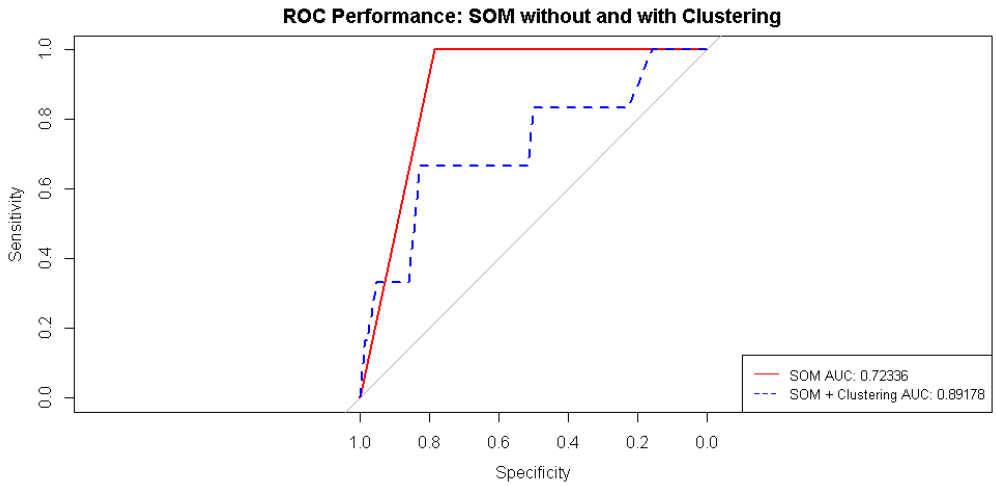


Figure 59. Clustering + SOM ROC.

Naïve Bayes Results

The naïve Bayes classifier produced probabilities for each CC code based on the probability of a readmit. Table 41 reports the highest probabilities of the readmit, with a CC code of 92 indicating a 20.1% probability. The naïve Bayes classifier significantly underperformed compared to the other analytics, in that the highest ROC achieved was 0.504 with the 50% training/50% validation model.

Table 42. Naive Bayes with Clustering Results.

CC92	**0**	**1**
Y		
0	0.812	0.187
1	0.789	0.201
CC96		
Y		
0	0.812	0.187
1	0.974	0.026
CC100		
Y		
0	0.814	0.185
1	0.978	0.021
CC108		
Y		
0	0.976	0.023
1	0.996	0.003

Naïve Bayes Validation

The naïve Bayes classifier was able to identify all non-readmits, but it was not able to quantify very many actual admissions, as evidenced from the 100% specificity and 0.08% sensitivity. This analytic performed the worst of each of the analytics with clustering membership.

Table 43. ROC Performance for Naive Bayes with Clustering.

	30Day Readmission	
Train%/Test%	Naïve Bayes ROC	Naïve Bayes + Clustering ROC
50%/50%	0.666	0.504
60%/40%	0.713	0.501
70%/30%	0.721	0.501
75%/25%	0.623	0.502
80%/20%	0.618	0.502

Table 44. Performance Comparison of Naive Bayes with Clustering.

Model	Predictive Accuracy Measures					
	Sensitivity	Specificity	Accuracy	PPV	NPV	AUROC
Naïve Bayes	0.444	0.998	0.993	0.571	0.996	0.721
Naïve Bayes + Clustering	0.008	1.000	0.910	1.00	0.916	0.504

Summary

By performing a cluster analysis on the data and segmenting the clusters into their own data sets, three of the four analytics were able to not only improve, but significantly improve the results. Additionally, by modeling readmissions as a function of condition categories as opposed to diagnoses columns, the models were able to expand current views of relationships of condition categories that predict readmissions.

Table 45. Summary of Best ROC per Analytic with Clustering.

Analytic	**Best ROC without Clusters**	**Best ROC with Clusters**	**Training Level**
Classification Decision Tree	**0.819**	**0.917**	**80%**
Neural Network	0.798	0.827	70%
Self Organizing Map	0.723	0.892	75%
Naïve Bayes	0.721	0.504	50%

Overall, the models that did benefit from clustering also showed more consistent prediction gains as the number of observations in the training set increased. Effectively, the models were better trained with more information while resisting overfit, with the exception being the naïve Bayes classifier.

Table 46. Summary of ROC Performance with Clustering.

	Decision Tree	Neural Net	SOM	Naïve Bayes
50%/50%	0.907	0.818	0.662	**0.504**
60%/40%	0.907	0.818	0.706	0.501
70%/30%	0.909	**0.827**	0.859	0.501
75%/25%	0.909	0.823	**0.892**	0.502
80%/20%	**0.917**	0.818	0.850	0.502

The next chapter discusses the implications and contributions of Chapters V and VI's results. Additionally, it discusses the future direction of the research and concludes the dissertation.

CHAPTER VII

DISCUSSION AND CONCLUSIONS

This chapter contains four elements. First, a discussion is provided around the advanced analytics performance. Next, the limitations of the research are provided. Future directions for the research are then discussed. The final section of this chapter provides a discussion around the conclusions to be drawn from this dissertation.

Discussion

This dissertation has examined the existing heart failure readmission prediction model provided by the Centers for Medicare and Medicaid Services. The provided model was applied to healthcare administrative data for the North Carolina State Health Plan. In so doing, a replication study has been conducted. The model's performance was in line with the performance described in the paper from which the model originated, with a highest predictive ability as measured by the ROC curve of 63.0%. This dissertation contributes in utilizing the model provided by CMS on state level data.

Additionally, the replication found statistical significance for four conditions in the NCSHP data: congestive heart failure, cardio-respiratory failure/shock, dialysis status, and major organ transplant status. The CMS logit model provides two views of the data: prediction ability, and the aggregate statistical significance of the factors that contribute

to the prediction ability. This replication established the baseline by which the advanced analytics must either expand, or improve.

Improvement, for the purposes of this dissertation is shown by providing a better prediction ability as measured by the ROC curve. Expansion, for the purposes of this dissertation, is shown by providing new views of the factors that contribute to readmissions. Readmissions are shown in Chapter III to be defined as both a quality measure, as well as have significant financial impact on hospitals. In so doing, the research questions driving this dissertation has been answered:

1. How can we improve and expand upon our understanding of the variances in the cost of care for specific health conditions?
2. How can we improve and expand upon our understanding of the variances in the quality of care for specific health conditions?

Specifically, to address research question one, this dissertation provides validation in the form of ROC curve statistics to each analytic that was tested. All of the analytics that were examined were able to improve prediction. Additionally, some of the analytics, by their nature, were able to also expand understanding of the factors that contribution to readmission. In answering research question two, this dissertation applied cluster analysis and re-modeled the data using readmissions as a function of exclusively condition categories. Chapter V's intent was to show improvement, Chapter VI shows expansion.

Answering these questions provided an exploration of various advanced analytic techniques has been conducted in Chapters V and VI. Predicting readmissions provided

the selection of five algorithms which are discussed in Chapter II literature review. These algorithms include: classification decision trees, artificial neural networks, self organizing maps, the naïve Bayes classifier, and two-step clustering. The first four algorithms were chosen based on their ability to classify and predict data. Clustering, being unable to predict, was used to expand understanding of readmission. Chapter V provides the results of four of the algorithms before applying clustering; Chapter VI provides the results of the four algorithms after applying clustering.

The results in Chapter V show that classification decision trees not only outperform the CMS model, but also outperform the other analytics. In so doing, classification decision trees have improved on understanding of how to predict heart failure readmissions. Additionally, classification decision trees expand upon knowledge of readmissions by providing decision heuristics in a manner that the logit model cannot. This is manifest by the sequencing of the decision nodes. The logit model provides statistical significance of conditions; classification decision trees provide statistical significance of decision splits based on the significance of conditions. In so doing, the classification decision tree analytic has both improved and expanded knowledge around heart failure readmits. Specifically, the NCSHP dataset's decisions provided a template to examining the profile of a patient at risk of a readmission. This template was relatively consistent throughout each training and validation level presented in Chapter V.

Decision trees for Chapter VI provided even further expansion of understanding of a profile of heart failure readmits. The two-step clustering algorithm was able to

separate non-readmits from readmits, with cluster seven being a cluster of readmission. Cluster seven contained a reduced set of conditions that were demonstrated to provide are more accuracy of predicting a readmit. By incorporating the decision heuristics generated from the decision tree with the classification from the two-step clustering, this dissertation has shown both improvement as well as expansion of heart failure readmissions.

The results of the artificial neural network analytic showed an improvement on prediction ability in Chapter V. Since the neural networks operate as a black box to prediction, the results presented in this dissertation focus on the improvement of predicting readmission, without expanding. The neural networks show potential to expand understanding of readmission, if the data is modeled in a different way, as is the case with Chapter VI. With the neural network results by using the clustering method, the pathways contain more information around each condition as opposed to simply providing information around the diagnoses columns.

The prediction performance of the self organizing map algorithm exceeded the baseline, but was underperformed by comparison with the other advanced analytics in this dissertation. The algorithm was selected based on its ability to take multidimensional data and reduce it into an x-y map of information. Self organizing maps expanded current understanding of readmission by providing visual cues to which factors in the data are aligned within each node. Additionally, the combination of those factors related between nodes is displayed.

The naïve Bayes classifier, by its own definition, is a classification algorithm that employs Bayes theorem to generate probabilities. Its performance excelled when using the full dataset in Chapter V, but it lost almost all prediction power after applying clustering. This result follows in line with the classifier assuming independence of assumptions in the data, and cluster analysis intentionally seeks to separate data dependencies. The classifier was able to improve prediction in Chapter V, while also expanding understanding of readmission by providing probabilities of readmits for each condition category code.

Overall, the classification decision tree algorithm was able to consistently improve and expand information around the problem of predicting heart failure readmits. Cluster analysis was able to improve the decision tree's ability. Cluster analysis was able to improve three of the four algorithm's prediction ability, effectively improving understanding of readmission, as well as model readmission in a different way, effectively expanding understanding of readmission.

Limitations and Future Research

The data used for this dissertation comes from the North Carolina State Health Plan, and represents 700,000 individuals subscribed to the health plan. As such, the majority of the data is representative of individuals living in the state of North Carolina.

Additionally, the data used for analysis was based on a subset of the population at the age of 65 or older. The data was subset because individuals at the age of 65 become

eligible for Medicare as their primary payer of insurance. This was by design so that inferences could be drawn upon a Medicare eligible population. This design also limits the conclusions to be drawn from the specific results to a population eligible for Medicare.

These limitations also present future research opportunities. While the data is confined to a population same from within a single state, there are opportunities to obtain similar healthcare administration data from other states and entities.

Additionally, only five algorithms were examined in this dissertation. These algorithms come from established literature without any of the various improvements that research has demonstrated in recent years. This creates a future research opportunity to employ state of the art algorithms to model the heart failure readmission problem, and improve and expand understanding of variances in heart failure readmissions.

There is also opportunity to pair this data with publically available census data. Advanced analytics benefit from having more data, and more dimensions of the data. In pairing the healthcare administrative data with information around the patient's zip code, for example, it can be tested which factors, if any help to improve the prediction ability.

Contributions and Conclusions

This dissertation contributes in several ways. First, a design science framework is established using a theoretical basis from literature to establish a process by which an investigation into applying advanced analytics to healthcare data can be conducted.

Second, this dissertation explicates this framework and search process through the extraction, transformation, and loading of general administrative healthcare data. Third, this dissertation contains a replication using state data of CMS sanctioned generalized logit model. Fourth, this dissertation compares the performance of four advanced analytic techniques at various training levels, and presents the optimal training level for each analytic, the prediction ability in terms of ROC curve, and the information generated from the analytic. Fifth, this dissertation presents the results of a cluster analysis and the ability of cluster analysis to improve prediction ability of advanced analytics.

This dissertation focused on one specific aspect of healthcare management, heart failure readmissions. It provides a framework for exploring other healthcare conditions using advanced analytics, in an effort to improve and expand knowledge in the healthcare domain.

REFERENCES

Abidi, S. S. R. (2001). Knowledge management in healthcare: Towards "knowledge-driven" decision-support services. *International Journal of Medical Informatics*, *63*(1-2), 5–18. doi:10.1016/S1386-5056(01)00167-8

Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *ACM SIGMOD Record*, *22*(2), 207–216. doi:10.1145/170036.170072

Agrawal, R., & Shim, K. (1996). Developing Tightly-Coupled Data Mining Applications on a Relational Database System. *LDD*, 287.

Agrawal, R., & Srikant, R. (1994). Fast Algorithms for Mining Association Rules. *Proceedings of the 20th International Conference on Very Large Data Bases*, 487–499. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.40.7506

AHRQ. (2013a). *Heart Failure Mortality Rate Technical Specifications IQI16* (Vol. 14).

AHRQ. (2013b). *Heart Failure Mortality Rate Technical Specifications PCI 08* (Vol. 14).

AHRQ. (2015a). *Death Rate among Surgical Inpatients with Serious Treatable Complications Technical Specifications PSI 04*.

AHRQ. (2015b). Inpatient Quality Indicators Overview. Retrieved October 13, 2015, from http://www.qualityindicators.ahrq.gov/Modules/iqi_resources.aspx

AHRQ. (2015c). Patient Safety Indicators Overview. Retrieved October 13, 2015, from http://www.qualityindicators.ahrq.gov/Modules/psi_resources.aspx

AHRQ. (2015d). Pediatric Quality Indicators Overvie. Retrieved October 13, 2015, from http://www.qualityindicators.ahrq.gov/Modules/pdi_resources.aspx

AHRQ. (2015e). Prevention Quality Indicators Overview. Retrieved October 13, 2015, from http://www.qualityindicators.ahrq.gov/Modules/pqi_resources.aspx

Altman, N. S. (1992). An Introduction to Kernel and Nearest-Neighbour Nonparametric Regression. *The American Statistician*, *46*(3), 175–185.

Andrienko, N., & Andrienko, G. (2013). Visual analytics of movement: An overview of methods, tools and procedures. *Information Visualization*, *12*(1), 3–24. doi:10.1177/1473871612457601

Ayodele, T. O. (2010). Machine Learning Overview. In *Machine learning overview* (pp. 9–19).

Bardhan, I. ., Oh, J.-H. ., Zheng, Z. ., & Kirksey, K. . (2015). Predictive analytics for readmission of patients with congestive heart failure. *Information Systems Research*, *26*(1), 19–39. doi:10.1287/isre.2014.0553

Bardhan, I. R., & Thouin, M. F. (2013). Health information technology and its impact on the quality and cost of healthcare delivery. *Decision Support Systems*, *55*(2), 438–449. doi:10.1016/j.dss.2012.10.003

Barnes, S., Hamrock, E., Toerper, M., Siddiqui, S., & Levin, S. (2015). Real-time prediction of inpatient length of stay for discharge prioritization. *Journal of the American Medical Informatics Association*, (301), 1–9. doi:10.1093/jamia/ocv106

Barton, D. (2012). Making Advanced Analytics Work For You. *Harvard Business Review*. Retrieved from http://www.buyukverienstitusu.com/s/1870/i/Making_Advanced_Analytics_Work_For_You.pdf

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Studies in the History of Statistics and Probability*, *1*, 134–153. doi:10.1016/B978-044450871-3/50096-6

Belmont, P. J., Garcia, E. J., Romano, D., Bader, J. O., Nelson, K. J., & Schoenfeld, A. J. (2014). Risk factors for complications and in-hospital mortality following hip fractures: a study using the National Trauma Data Bank. *Archives of Orthopaedic and Trauma Surgery*, *134*(5), 597–604. doi:10.1007/s00402-014-1959-y

Berwick, D. M., Nolan, T. W., & Whittington, J. (2008). The triple aim: Care, health, and cost. *Health Affairs*, *27*(3), 759–769. doi:10.1377/hlthaff.27.3.759

Breault, J. L., Goodall, C. R., & Fos, P. J. (2002). Data mining a diabetic data warehouse. *Artificial Intelligence in Medicine*, *26*(1-2), 37–54. doi:10.1016/S0933-3657(02)00051-9

Britton, M. (2003). The burden of COPD in the U.K.: results from the confronting COPD survey. *Respiratory Medicine*, *97*, S71–S79. doi:10.1016/S0954-6111(03)80027-6

Brownlee, J. (2015). A Tour of Machine Learning Algorithms. Retrieved October 13, 2015, from http://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/

Burkey, M., Bhadury, J., & Eiselt, H. A. (2011). Voronoi Diagrams and Their Uses. In *Journal of Chemical Information and Modeling* (pp. 445–470). Springer US. doi:10.1017/CBO9781107415324.004

Calvillo–King, L., Arnold, D., Eubank, K. J., Lo, M., Yunyongying, P., Stieglitz, H., & Halm, E. A. (2013). Impact of Social Factors on Risk of Readmission or Mortality in Pneumonia and Heart Failure: Systematic Review. *Journal of General Internal Medicine*, *28*(2), 269–282. doi:10.1007/s11606-012-2235-x

Carbonell, J., Michalski, R., & Mitchell, T. (1983). An Overview of Machine Learning. *Machine Learning SE - 1*. doi:10.1007/978-3-662-12405-5_1

Centers for Medicare and Medicaid Services. (2016, January). Risk-Adjustors.

Chae, Y. M., Kim, H. S., Tark, K. C., Park, H. J., & Ho, S. H. (2003). Analysis of healthcare quality indicator using data mining and decision support system. *Expert Systems with Applications*, *24*(2), 167–172. doi:10.1016/S0957-4174(02)00139-2

Chan, C. M. L. (2013). From Open Data to Open Innovation Strategies: Creating E-Services Using Open Government Data. *2013 46th Hawaii International Conference on System Sciences*, 1890–1899. doi:10.1109/HICSS.2013.236

Chen, H., Chiang, R., & Storey, V. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly*, *36*(4), 1165–1188.

Cheng, J., Bell, D. a., & Liu, W. (1997). An Algorithm for Bayesian Belief Network Construction from Data. *Proceedings of the Sixth International Workshop on Artificial Intelligence and Statistics (AI & STAT '97)*, 83–90.

Chesbrough, H. W., & Appleyard, M. M. (2007a). Open innovation and strategy. *California Management Review*, *50*(1), 57–+. doi:10.1016/j.jbiosc.2010.11.012

Chung, W., Chen, H., & Jr, J. N. (2005). A visual framework for knowledge discovery on the Web: An empirical study of business intelligence exploration. *Journal of Management Information Systems*, *21*(4), 57–84.

Churilov, L., Bagirov, A., Schwartz, D., Smith, K., & Dally, M. (2005). Data Mining with Combined Use of Optimization Techniques and Self-Organizing Maps for

Improving Risk Grouping Rules: Application to Prostate Cancer Patients. *Journal of Management Information Systems*, *21*(4), 85–100. doi:Article

CMS. (2015a). *NHE Fact Sheet, 2013*. Retrieved from https://www.cms.gov/research-statistics-data-and-systems/statistics-trends-and-reports/nationalhealthexpenddata/nhe-fact-sheet.html

CMS. (2015b, April 17). Quality Measures. Retrieved September 24, 2015, from https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/QualityMeasures/index.html?redirect=/qualitymeasures/03_electronicspecifications.asp

Dasgupta, A., Sun, Y. V., König, I. R., Bailey-Wilson, J. E., & Malley, J. D. (2011). Brief review of regression-based and machine learning methods in genetic epidemiology: the Genetic Analysis Workshop 17 experience. *Genetic Epidemiology*, *35*(S1), S5–S11. doi:10.1002/gepi.20642

Davenport, T. (2013). Analytics 3.0: in the new era, big data will power consumers products and services. *Harvard Business Review*, (December 2013).

Dempster, a. P. A., Laird, N. M. N., & Rubin, D. D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B Methodological*, *39*(1), 1–38. doi:http://dx.doi.org/10.2307/2984875

Dharmarajan, K., Hsieh, A. F., Lin, Z., Bueno, H., Ross, J. S., Horwitz, L. I., … Krumholz, H. M. (2013). Diagnoses and timing of 30-day readmissions after hospitalization for heart failure, acute myocardial infarction, or pneumonia. *Jama*, *309*(4), 355–63. doi:10.1001/jama.2012.216476

E1071), TU Wien. R package version 1.6-7. http://CRAN.R-project.org/package=e1071

El Mensouri, R., Beqali El, O., & Elhoussaine, Z. (2013). Blind Optimization for data Warehouse during design. *International Journal of Computer Science and Information Security*, *11*(10), 27–32.

Fang, X., Hu, P. J., Li, Z. L., & Tsai, W. (n.d.). Predicting Adoption Probabilities in Social Networks Predicting Adoption Probabilities in Social Networks, (April 2014).

Farquhar, M. (2008). Chapter 45 . AHRQ Quality Indicators. *Patient Safety and Quality: An Evidence-Based Handbook for Nurses*, 40.

Fayyad, U. M. (1996). Data mining and knowledge discovery: Making sense out of data. *IEEE Intelligent Systems*, *11*(5), 20–25.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases, *17*(3), 37. doi:10.1609/aimag.v17i3.1230

Fischer, C., Steyerberg, E. W., Fonarow, G. C., Ganiats, T. G., & Lingsma, H. F. (2015). A systematic review and meta-analysis on the association between quality of hospital care and readmission rates in heart failure patients. *American Heart Journal*, *170*(5), 1005–1017.e2. doi:10.1016/j.ahj.2015.06.026

Graves, N., Halton, K., Doidge, S., Clements, a., Lairson, D., & Whitby, M. (2008). Who bears the cost of healthcare-acquired surgical site infection? *Journal of Hospital Infection*, *69*(3), 274–282. doi:10.1016/j.jhin.2008.04.022

Gregor, S. (2006). The nature of theory in information systems. *Mis Quarterly*, *30*(3), 611–642.

Gregor, S., & Hevner, A. R. (2013). Positioning and Presenting Design Science Research for Maximum Impact. *MIS Quarterly*, *37*(2), 337–355. doi:10.2753/MIS0742-1222240302

Guzella, T. S., & Caminhas, W. M. (2009). A review of machine learning approaches to Spam filtering. *Expert Systems with Applications*, *36*(7), 10206–10222. doi:10.1016/j.eswa.2009.02.037

Halfon, P., Eggli, Y., Prêtre-Rohrbach, I., Meylan, D., Marazzi, A., & Burnand, B. (2006). Validation of the potentially avoidable hospital readmission rate as a routine indicator of the quality of hospital care. *Medical Care*, *44*(11), 972–981. doi:10.1097/01.mlr.0000228002.43688.c2

Hansen, M. M., Miron-Shatz, T., Lau, a Y. S., & Paton, C. (2014). Big Data in Science and Healthcare: A Review of Recent Literature and Perspectives. Contribution of the IMIA Social Media Working Group. *Yearbook of Medical Informatics*, *9*(1), 21–6. doi:10.15265/IY-2014-0004

Hey, T. (2010). The next scientific revolution. *Harvard Business Review*, *88*(11), 56–63, 150. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/21049680

Hothorn, T., Hornik, K. & Zeileis, A. (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. Journal of Computational and Graphical Statistics, 15(3), 651--674.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, *31*(8), 651–666. doi:10.1016/j.patrec.2009.09.011

Jain, A., Murty, M., & Flynn, P. (1999). Data clustering: a review. *ACM Computing Surveys (CSUR)*.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, *32*(3), 241–254. doi:10.1007/BF02289588

Kansagara, D., Englander, H., Salanitro, A., Kagen, D., Theobald, C., Freeman, M., & Kripalani, S. (2011). Risk Prediction Models for Hospital Readmission: A Systematic Review. *Journal of American Medical Association*, *306*(15), 1688–1698.

Kelarev, A., Abawajy, J., Stranieri, A., & Jelinek, H. (2013). Empirical Investigation of Decision Tree Ensembles for Monitoring Cardiac Complicatons of Diabetes. *International Journal of Data Warehousing and Mining*, 1–18.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, *43*(1), 59–69. doi:10.1007/BF00337288

Kotsiantis, S., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques, *31*, 249–268.

Krumholz, H., Normand, S., Keenan, P., Lin, Z., Drye, E., Bhat, K., … Schreiner, G. (2008). *Hospital 30-Day Heart Failure Readmission Measure*.

Kumar, A., Niu, F., & Ré, C. (2013). Hazy: Making it Easier to Build and Maintain Big-Data Analytics. *Communications of the ACM*, *56*(3), 40. doi:10.1145/2428556.2428570

Lakoumentas, J., Drakos, J., Karakantza, M., Sakellaropoulos, G., Megalooikonomou, V., & Nikiforidis, G. (2012). Optimizations of the naïve-Bayes classifier for the prognosis of B-Chronic Lymphocytic Leukemia incorporating flow cytometry data. *Computer Methods and Programs in Biomedicine*, *108*(1), 158–167. doi:10.1016/j.cmpb.2012.02.009

Langfelder, P., & Horvath, S. (2012). Fast R Functions for Robust Correlations and Hierarchical Clustering. *Journal of Statistical Software*, *46*(11). doi:i11 [pii]

Lavalle, S., Lesser, E., Shockley, R., Hopkins, M. S., & Kruschwitz, N. (2011). Big Data, Analytics and the Path From Insights to Value. *MIT Sloan Management Review*, *52*(2), 21–32.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. doi:10.1038/nature14539

Lefèvre, T., Rondet, C., Parizot, I., & Chauvin, P. (2014). Applying Multivariate Clustering Techniques to Health Data: The 4 Types of Healthcare Utilization in the Paris Metropolitan Area. *PLoS ONE*, *9*(12), e115064. doi:10.1371/journal.pone.0115064

Ma, C.-T. A. (1994). Health Care Payment Systems-Cost and Quality Incentives.Pdf. *Journal of Economics & Management Strategy*.

MacQueen, J. B. (1967). Kmeans Some Methods for classification and Analysis of Multivariate Observations. *5th Berkeley Symposium on Mathematical Statistics and Probability 1967*, *1*(233), 281–297. doi:citeulike-article-id:6083430

March, S., & Smith, G. (1995a). Design and Natural Science Research on Information Technology. *Decision Support Systems*, *15*, 251–266. doi:10.1016/0167-9236(94)00041-2

March, S., & Smith, G. (1995b). Design and Natural Science Research on Information Technology. *Decision Support Systems*, *15*, 251–266. doi:10.1016/0167-9236(94)00041-2

Matwin, S., & Sazonova, V. (2012). Direct comparison between support vector machine and multinomial naive Bayes algorithms for medical abstract classification. *Journal of the American Medical Informatics Association*, *19*(5), 917–917. doi:10.1136/amiajnl-2012-001072

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2015). e1071: Misc Functions of the Department of Statistics, Probability Theory Group

Mullins, I. M., Siadaty, M. S., Lyman, J., Scully, K., Garrett, C. T., Greg Miller, W., … Knaus, W. a. (2006). Data mining and clinical data repositories: Insights from a 667,000 patient data set. *Computers in Biology and Medicine*, *36*(12), 1351–1377. doi:10.1016/j.compbiomed.2005.08.003

NBER. (2016). ICD HCC Crosswalk -- International Classification of Diseases to Hierarchical Condition Category Crosswalk.

Niu, F., Zhang, C., Ré, C., & Shavlik, J. (2012). Elementary: Large-scale Knowledge-base Construction via Machine Learning and Statistical Inference. *International Journal On SemanticWeb and Information Systems*, (Special Issue onWeb-Scale Knowledge Extraction). doi:10.4018/jswis.2012070103

Patient Protection and Affordable Care Act (2010). United States of America.

Phillips-wren, G., Iyer, L., Kulkarni, U., & Ariyachandra, T. (2015). Business Analytics in the Context of Big Data: A Roadmap for Research Roadmap for Research. *Communications of the AIS*, *37*.

Phillips-Wren, G., Sharkey, P., & Dy, S. M. (2008). Mining lung cancer patient data to assess healthcare resource utilization. *Expert Systems with Applications*, *35*(4), 1611–1619. doi:10.1016/j.eswa.2007.08.076

Pope, G., Ellis, R., Ash, A., Ayanian, J., Bates, D., Burstin, H., … Wu, B. (2000). *Diagnostic Cost Group Hierarchical Condition Category Models for Medicare Risk Adjustment*.

Public Law 104-191. Health Insurance Portability and Accountability Act of 1996 (1996).

Quinlan, J. R. (1986). Induction of Decision Trees. *Machine Learning*, *1*, 81–106.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. *Morgan Kaufmann Publishers*.

Rahimi, K., Bennett, D., Conrad, N., Williams, T. M., Basu, J., Dwight, J., … MacMahon, S. (2014). Risk Prediction in Patients With Heart Failure. *JACC: Heart Failure*, *2*(5), 440–446. doi:10.1016/j.jchf.2014.04.008

Rana, S., Gupta, S., Phung, D., & Venkatesh, S. (2015). A Predictive Framework for Modeling Healthcare Data with Evolving Clinical Interventions. *Statistical Analysis and Data Mining*, *8*(3), 162–182. doi:10.1002/sam

Rish, I. (2001). An empirical study of the naive Bayes classifier. *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, *22230*, 41–46.

Rodolfo, a, Pérez-ortega, J., Miranda-henriques, F., & Reyes-salgado, G. (2010). Spatial Data Mining of a Population-Based Data Warehouse of Cancer in Mexico. *International Journal of Combinatorial Optimization Problems and Informatics*, *1*(1), 61–67.

Rumelhart, D., Hinton, G., & Williams, R. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536.

Serra-Sutton, V., Serrano, C. B., & Carreras, M. E. (2013). Quality Indicators To Assess a Colorectal Cancer Prevention Program. *International Journal of Technology Assessment in Health Care*, *29*(02), 166–173. doi:10.1017/S0266462313000020

Shmueli, G., & Koppius, O. (2011). Predictive Analytics in Information Systems Research. *MIS Quarterly*, (September), 553–573.

Sokol, M. C., McGuigan, K. a, Verbrugge, R. R., & Epstein, R. S. (2005). Impact of medication adherence on hospitalization risk and healthcare cost. *Medical Care*, *43*(6), 521–530. doi:10.1097/01.mlr.0000163641.86870.af

Speybroeck, N. (2012). Classification and regression trees. *International Journal of Public Health*, *57*(1), 243–246. doi:10.1007/s00038-011-0315-z

Stelfox, H. T., Straus, S. E., Nathens, A., & Bobranska-Artiuch, B. (2011). Evidence for quality indicators to evaluate adult trauma care: A systematic review*. *Critical Care Medicine*, *39*(4), 846–859. doi:10.1097/CCM.0b013e31820a859a

Takeda, S., Taylor, S., Taylor, R., Kahn, F., Krum, H., & Underwood, M. (2012). Clinical service organisation for heart failure. *Cochrane Database of Systematic Reviews (Online)*, *12*(9), CD002752. doi:10.1002/14651858.CD002752.pub3

The White House. (2013). *Economic Report of the President*.

Tsai, A. ., Williamson, D., & Glick, H. (2011). Direct medical cost of overweight and obesity in the United States a quantitative systematic review. *International Association for the Study of Obesity*, *12*(1), 50–61. doi:10.1111/j.1467-789X.2009.00708.x.Direct

Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

R. Wehrens and L.M.C. Buydens, Self- and Super-organising Maps in R: the kohonen package J. Stat. Softw., 21(5), 2007

Walls, J. G., Widmeyer, G., & Sawy, O. a E. (1992). Building an Information System Design Theory for Vigilant EIS. *Information Systems Research*, *3*(1), 36–60.

Wang, X., Zheng, B., Good, W., King, J., & Chang, Y. (1999). Computer-assisted diagnosis of breast cancer using a data-driven Bayesian belief network. *International Journal of Medical Informatics*, *54*(2), 115–126. doi:10.1016/S1386-5056(98)00174-9

Weisbrod, B. (1991). THe Health Care Quadrilemma : Essay on Technological Change , of Care , Quality and Insurance , Cost Containment. *Journal of Economic Literature*, *29*(2), 523–552.

Wongravee, K., Lloyd, G. R., Silwood, C. J., Grootveld, M., & Brereton, R. G. (2010). Supervised self organizing maps for classification and determination of potentially discriminatory variables: illustrated by application to nuclear magnetic resonance metabolomic profiling. *Analytical Chemistry*, *82*(2), 628–38. doi:10.1021/ac9020566

Zaki, M. J. (2000). Scalable algorithms for association mining. *IEEE Transactions on Knowledge and Data Engineering*, *12*(3), 372–390. doi:10.1109/69.846291

Zaki, M. J. M., Parthasarathy, S., Ogihara, M., Li, W., & Others. (1997). New algorithms for fast discovery of association rules. *Kdd*, *7*, 283–286.

Zhang, J. X., Rathouz, P. J., & Chin, M. H. (2003). Comorbidity and the concentration of healthcare expenditures in older patients with heart failure. *Journal of the American Geriatrics Society*, *51*(4), 476–482. doi:10.1046/j.1532-5415.2003.51155.x

Zhu, B., & Watts, S. a. (2010). Visualization of Network Concepts: The Impact of Working Memory Capacity Differences. *Information Systems Research*, *21*(2), 327–344. doi:10.1287/isre.1080.0215

Zhu, X. (2007). Semi-Supervised Learning Literature Survey. *Sciences-New York*, 1–59. doi:10.1.1.146.2352