The Untold Story of Predicting Readmissions for Heart Failure Patients

Ahmad K. Aljadaan

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2019

Reading Committee:

John H. Gennari, Chair

Peter Tarczy-Hornoch

Adam Wilcox

Program Authorized to Offer Degree:

Biomedical and Health Informatics

University of Washington

Abstract

The Untold Story of Predicting Readmissions for Heart Failure Patients

Ahmad Khalid Aljadaan

Chair of the Supervisory Committee:
John H. Gennari
Biomedical Informatics and Medical Education

The availability and accessibility of Electronic Health Record (EHR) data create an opportunity for researchers to revolutionize healthcare. The recognition of the importance of secondary use of EHR data has led to the development of research-ready integrated data repositories (IDRs) from EHR data. Analyzing this data can help researchers connect the dots and can lead to critical clinical findings through predictive analytics methods. Unfortunately, poor data quality is a problem that affects the accuracy of such findings. An example of a data quality problem is poor information about the specifics of admission, discharge, and readmission.

Heart Failure (HF) is one of the most common cardiovascular diseases. 5.7 million people in the United States have heart failure with 870,000 new cases annually, and this disease is the leading cause of hospital readmission.

Predicting readmission for heart failure patients has been well-studied. The readmission periods that researchers have studied range between 30 days to one year. However, shorter than 30 days readmission have received less research attention. In my research, I shed light on an overlooked yet important group of readmissions: very early readmissions. Currently, little is known about what causes heart failure patients to come back so quickly. In the long term, my career goal is to predict very early readmission patients before discharge and improve on the discharge decision making. It is a step toward personalized healthcare to improve patient care eventually.

The broad goal of my dissertation is to leverage the availability and accessibility of electronic health data and characterize day 1-30 readmission, more specifically characterizing very early readmissions. My approach to reach my goal went through four major steps: 1) Reviewing the literature to understand the field and how early readmission have been defined, 2) Using retrospective EHR data from UW Medicine to build an accurate visit table for heart failure patients, 3) Using the visit table to build a prediction model to characterize day 1-30 readmissions, 4) Improving on the model by applying different machine learning algorithms and imputation techniques for missing data.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I want to express my sincere gratitude and appreciation to my mentor and advisor during my Ph.D. journey, John Gennari. It was a five-year full of support and encouragement during the ups and downs of this journey. I believe that I was lucky enough to experience the gold standard of advising. John unlimited support, guidance, patient, and mentorship made me a better researcher, and writer. I would also like to thank my committee members Peter Tarczy-Hornoch, Adam Wilcox, Annie Chen, and my GSR Shuai Huang for their great support and insights. Special thanks to committee member Todd Dardas for making this possible through his guidance as a Cardiologist and translating the domain knowledge into research language. Also acknowledged is the data support provided by the Health Resources Services Administration (HRSA) grant # UD7HP26909 and Principle Investigator Dr. Zierler.

Finally, special thanks to my family and friends, especially my parents, my brother, my sisters, my wife, my children; Khalid, Abdullah, and Sarah, and my Biomedical and Health Informatics friends. The love, support, and encouragement they provided me have been the fuel that pushed me to become who I am.

DEDICATION

To my mother and father, who passionately supported me in every possible way and encouraged me to be the best version of myself. To my beautiful wife, Maram Alsoghayyer- who without her, I have not accomplished what I have accomplished.

Chapter 1. INTRODUCTION AND ROADMAP

The availability and accessibility of Electronic Health Record (EHR) data create an opportunity for researchers to revolutionize healthcare. The recognition of the importance of secondary use of EHR data has led to the development of research-ready integrated data repositories (IDRs) from EHR data. Analyzing this data can help researchers connect the dots and can lead to important clinical findings through predictive analytics methods. Unfortunately, poor data quality is a problem that affects the accuracy of such findings. An example of a data quality problem is poor information about the specifics of admission, discharge, and readmission.

In my dissertation, I would like to leverage the availability of data to understand the readmission patterns for heart failure patients. Giving the data quality issues that exist in IDRs, I will start by creating a visit table that accurately captures temporal information about all admissions and discharges for patients. This is not a trivial problem for researchers as noise do exist in the data due to billing procedures, human error, or some other reasons affecting the accuracy of the readmission patterns. Therefore, it is important to go through rigorous data cleaning and validation from the identified EHR data, discharge summary report, and manual chart review for some patients.

I chose to focus on heart failure patients because of its epidemic, high risk of mortality, and high treatment cost. More specifically, I will be analyzing “day 1-30 readmission” visits to find subgroups (clusters) for 30 readmissions and find the parameters that could help predict very early readmission (earlier than 30 days). I want to locate the range of days when very early readmission occurs. I hypothesize that very early readmission are those readmissions that are more likely due to error, or readmissions, that if avoided, could have led to substantial cost savings and/or potentially improved health outcomes. The goal of my research is to predict which patients might become members of "very early readmission" patients at the time of discharge to help personalize the decision of discharge for this group. For these reasons, I want to be able to highlight their visit characteristics and help clinicians' flag these visits before they are discharged to avoid unnecessary very early readmissions.

1.1 MOTIVATION FOR RESEARCH

There are two driving forces behind this work. First, Heart Failure (HF) is one of the most common cardiovascular diseases where 5.7 million people in the United States have heart failure with 870,000 new cases annually. This number is projected to increase by 46% in 2030 to reach 8 million people. Currently, 1 in 9 death certificates in the United States mentioned heart failure. (Benjamin et al., 2018) It is the leading cause of hospital readmission where 25% of patients of Medicare beneficiaries are readmitted within 30 days. (Kheirbek et al., 2015) (Hersh, Masoudi, & Allen, 2013) Heart failure readmission is the most frequent cause of potentially avoidable readmission. (Donzé, Lipsitz, David, & Jeffrey, 2013) Moreover, the cost of care for the heart failure patient is estimated to be around $39.2 billion annually. (Voigt et al., 2014) The Affordable Care Act (ACA) third stage started to penalize hospitals with readmission rate higher than the proposed excess readmission ratio. The readmission ratio is built to help calculate the excess readmission ratios, which is a measure of a hospital readmission performance compared to the national average. ("Hospital Readmissions Reduction Program (HRRP)," 2018) Thus, the need to reduce hospital readmission for congestive heart failure is becoming vital for both reasons; improving healthcare and saving costs.

Second, the availability of EHR data gives researchers an excellent opportunity to help reduce the epidemic of heart failure and its associated costs. For that reason, there has been much research on predicting readmission and trying to minimize patient readmission within 30 days of discharge. (Ross et al., 2008)(O'Connor et al., 2016) However, shorter than 30 days readmissions ranges have received less research attention, which is the theme of my research. I want to investigate day 1-30 readmissions and analyze the differences in their characteristics with the aim of defining at what range of day do very early readmissions occur (shorter than 30 days). Also, finding the clusters that within 30 days where they share similar characteristics.

To my knowledge, there has been no prior research that characterized day 1 to 30 readmission and defined who is a rapid (or very early) readmission patient with heart failure disease and what patient characteristics could lead to very early readmission. Gabayan et al. defined rapid readmissions to the emergency department to be within seven days. However, this work is not specific for heart failure patients. (Gabayan et al., 2013) Moreover, Eastwood et al. have looked up the determinant factors that would increase the likelihood of early readmission for heart failure patients, focusing mainly on 7- and 30-days readmission. (Eastwood et al., 2014) Similar to Gayan et al. who looked into rapid emergency room readmission, they selected seven days because of an analysis that was done by Clarke in 1990 where he assigned 0-6 days avoidable readmission in general not specific for heart failure. (Clarke, 1990) An important outcome of my research is to define the range of days for very early readmission and understand the visit characteristics that help predict it. Also, I intend to locate the classifications in readmission ranges, if exist, among 30 days readmission visits.

Currently, little is known about what causes some heart failure patients to come back so quickly. The proposed model to predict very early readmission patients before discharge is a step toward personalized health care to improve patient care. I hypothesize that my model can help reduce very early readmissions and potentially lower the risk of mortality of HF patients and reduce the treatment cost.

1.2 SOLUTION APPROACH AND SCOPE

In order to best utilize the secondary usage data and shed light on an essential group of heart failure readmissions, I propose a definition of what is considered very early readmission. First, I need to understand the field of predicting heart failure readmission and what has been done in the literature. Although the literature has exhaustively investigated predicting 30 days readmission, none have characterized 30 days readmission and define very early readmission visits. In Chapter 2, I conduct a systematic review to see what has been done in predicting shorter than 30 days readmission, which I will talk about in section 1.2.1. Second, the process to build an accurate prediction model need a solid base represented by an accurate visit table. Giving the data quality issues that exist in the integrated data repositories (IDRs), I will start by creating a visit table that accurately captures temporal information about all admissions and discharges for patients and validates its accuracy, as described in section 1.2.2. After creating the heart failure visit table, I will start characterizing day 1-30 readmissions to define the range of days where very early readmission occurs, which I describe in section 1.2.3. Finally, I will try to improve upon the prediction model from the previous chapter by applying different machine learning algorithms and applying imputation technique to missing data, as described below in section 1.2.4.

1.2.1 *Predicting Readmission, A review of related work*

In chapter 2, I have researched prediction models on heart failure readmissions in the literature to understand the past and the present of the field. While there is a plethora of research about heart failure prediction models, I started my literature review with two seminal systematic reviews by

Ross et al. and O’Connor et al. (Ross et al., 2008) (O’Connor et al., 2016) Ross et al. looked into 941 articles from 1950 to 2007 that studied the risk of readmission for heart failure patients. Their study yielded with only five papers that met their inclusion and exclusion criteria. Ross et al. concluded from analyzing the five papers that no consistent predictors appeared from their review. The readmission ranges among the papers in the systematic review varies between 30 days to one year. There was no paper with shorter than 30 days readmission included. (Voigt et al., 2014) One reason for the focus on 30-day readmission is because the Affordable Care Act (ACA) third stage started to penalize hospitals with less than 30 days readmission rate higher than the proposed excess readmission ratio. A more recent systematic review by O’Connor et al., which searched the literature for studies that identified heart failure patient characteristics measured before discharge from 1992 to 2014. O’Connor et al. yielded a similar result as Ross et al. with not finding a single patient characteristic considered as the key factor for readmission among the different studies. From the 905 articles, O’Connor et al. included only one paper that analyzed shorter than 30 days readmission. Moreover, I have analyzed the studies that have been done on two registries, the Acute Decompensated Heart Failure National Registry (ADHERE) and the Organized Program to Initiate Lifesaving Treatment in Hospitalized Patients with Heart Failure (OPTIMIZE-HF) registries, which were created to bridge the gap between knowledge and medical care through the use of evidence-based research. Both studies on these two registries have not investigated shorter than 30 days readmission.

After investigating the current work on the literature, I conducted an updated and more specific review of the literature seeking short readmissions. I searched PubMed from 2014 to 2017 with the following MeSH Terms; ([“Heart Failure” OR “Cardiac Failure” OR “Congestive Heart Failure” OR “Heart Decompensation” OR “Heart Failure, Congestive” OR “Heart Failure, Left-

Sided" OR "Heart Failure, Right Sided" OR "Left-Sided Heart Failure" OR "Myocardial Failure" OR "Right-Sided Heart Failure"] AND ["Patient Readmission" OR "Hospital Readmission" OR "Hospital Readmissions" OR "Readmission, Hospital" OR "Readmissions, Hospital"] AND ["Hospitalization"]). Then, I added these Mesh Terms; ["Short Readmissions" OR "Rapid Readmissions" OR "Avoidable Readmissions" OR "Unplanned Readmissions" OR "Early Readmissions"]. "]. This yielded very few abstracts that are related to my research. Most abstracts refer to early, short, or avoidable readmission to be within 30 days. This new literature review confirms the early work that little work has been done to characterize shorter than 30 days readmission for heart failure patients.

1.2.2 *Counting Readmissions, Surprisingly Difficult*

In Chapter 3, I lay the foundation for my research by building a visit table for heart failure patients. To understand readmission patterns, I need an accurate visit table with patient admission and discharge data which can then be used to query for individuals' readmission history, or for institution-wide patterns. Unfortunately, I found that building such a visit table is surprisingly difficult. If noise or inaccuracies exist in this table, that will affect the accuracy of any predictive analytics that might be built from this information.

First, in consultation with a cardiologist; Dr. Todd Dardas, we defined the heart failure cohort using a precise definition. The definition consists of International Classification Code (ICD) and Diagnosis Related Groups (DRG) codes. Then, in consultation with cardiologists and via manual chart review, I developed a set of rules to help "clean" the visit table. These rules successively discard data that would be inappropriate to include for analysis of readmission patterns. For example, if a patient left against medical advice, and later is readmitted, that readmission should not count for analysis, if the aim is to build a system to improve physician discharge decisions.

The methods also included removing data that were marked as discharge but should be more appropriately viewed as a transfer. For example, sometimes a patient may be transferred between units within the hospital, yet these transfers are documented as a discharge, followed by readmission. Again, this sort of readmission should not be included in a visit if the goal is improved decision support.

These methods resulted in approximately a 6% reduction in data. After cleaning, I assessed the accuracy of the visit table via consultation with cardiology and manual chart review. Although not a very large percentage of data, this cleaning step could be critical to analysis, as it may be that the data cleaned are outliers, and might, therefore, have a significant impact on aggregate summaries that would be important for analyses.

1.2.3 *Characterizing Day 1-30 Readmissions*

In Chapter 4, I use the visit table created from chapter 3 to characterize 30 days readmission visits for heart failure patients to find subgroups for shorter than 30 days readmission with the focus on defining very early readmission characteristics. However, not all the predictor variables needed for the model currently exist in the visit table. For that, I started the data preparation for some variables and adding them to the visit table such as Medical History, In Hospital Medication, and Echocardiography data. During this process, I used natural language processing techniques to extract some variables from free text data. Also, some data needed to be cleaned and classified accurately such as Race and in-hospital Medication. Data cleaning has also been part of this work when dealing with missing data such as the Age variable.

After the data preparation, and the visit table needed for the prediction model is completed, I started setting up the model. Giving the precise definition of the cohort, the visit table yielded 1385 patients and 7194 visits. However, since my study focuses mainly on readmissions, I dropped the initial patient visit. This result in 1123 patients and 5488 visits. From those, 2030 visits from 688 heart failure patients happened within 30 days (~%37). Then, I carried on to build a Logistic Regression model with R that predicts if a patient will be readmitted within the predefined readmission range.

As proof of concept, I started by comparing my model with the Adhere model, using the same variables. Of the 46 variables evaluated for the risk of 30 readmissions, Systolic Blood Pressure at admission is the most predictable variable, and the C-Statistic score of the model is 0.64. Then, I added to the adhere variables the following variables; vital at discharge, In-hospital medication, Length of Stay, and Number of Previous Readmissions. The most significant variable of this model is Medicaid insurance, where the C-Statistic score of the model is 0.65. Finally, adding Echocardiography data to the model has slightly improved the C-Statistic score where the most predictable variable is the number of previous readmissions. Echocardiography or echo test is a test that takes “moving pictures” of the heart with sound waves. (American Heart Association., 2013) The test provides data about heart structure and performance.

In my analysis, I carried on with the third model since it has the highest prediction accuracy. I started modeling for 6-days ranges of readmission with the aim of finding the cut off when the patient characteristics started to change. Results of the logistic regression models among the six days ranges show that the variables Diuretics, Angiotensin Receptor Blocker, and Abnormal Echo are predictable variables for 1-6, and 2-7 days range of readmission. After the seventh day of readmission, the visit characteristics started to change at 3-8 days of readmission. There are 852 visits of the 2030 30-day readmission visits (~38%) occur in the first seventh days. Then, I started to see the similarity in visit characteristics for the ranges 8-13, and 9-15 days until readmission. The common, predictable variables for these ranges are Brain natriuretic peptide

(BNP) test at admission, Diabetes, and the number of previous readmissions. The third group of visits characteristics happens in the ranges 16-27 days readmissions with different visits characteristics from the previous two groups. The conclusion is that very early readmission happens between day 1-7 of readmission. The 1-7 days prediction model has C-Statistic score of 0.72 and Diuretics, Angiotensin Receptor Blocker, and abnormal echo test among the most predictable variable.

1.2.4 *Improving on the Model*

In Chapter 5, I worked on improving the accuracy of the model I built in chapter 4 by applying different machine algorithms and imputation technique for missing data. The goal of this chapter is to see if different machine algorithm and/or implementing imputation techniques will improve the prediction accuracy or change the classification of 30 days readmission by providing a different definition for very early readmission. For that, I selected Random Forest and Regularized Logistic regression (LASSO) algorithms for modeling. In prior research, these methods did not improve prediction scores for the 30-day readmission task, (Futoma, Morris, & Lucas, 2015) (Yang, Delcher, Shenkman, & Ranka, 2016) (Garcia-Arce, Rico, & Zayas-Castro, 2017)(Golas et al., 2018) but because my task is a bit different, I felt it essential to assess alternative prediction algorithms. I started by applying the Random Forest algorithm using the same variables used for the logistic regressions model in chapter 4. When applying the Random Forest algorithm, it yielded a 0.68 C-Statistic score with 0.88 sensitivity and 0.22 specificity. Random Forest algorithm has slightly better prediction accuracy than the logistics regression algorithm. Creatinine at admission, Systolic Blood Pressure at admission, and the number of previous readmissions are the most critical variables based on mean decrease Gini. Mean decrease Gini gives ranking scores of the variables, the larger the score, the more importance of the model. The mean decrease in Gini coefficient is a measure of how each variable contributes to the homogeneity of the nodes and leaves in the resulting random forest. (Menze et al., 2009) Then, I have applied LASSO Regression to check the accuracy of predicting 30 days readmission. Since LASSO cannot handle missing data, I needed to apply some strategy to fix the missingness of the data. For that, I applied the imputation technique with the assumption that the missingness is not at random (MNAR). (Wells, Nowacki, Chagin, & Kattan, 2013) (Beaulieu-Jones et al., 2018) The prediction accuracy of applying the LASSO algorithm after applying the MNAR imputation is 0.63. The most predictable variables are Hemoglobin at admission, the number of previous readmissions, and age.

Finally, I have applied the imputation on both Random Forest and Logistic regression to see if that will improve the prediction accuracy. For Logistic Regression, the C-Statistic score is 0.64 lower than the model without imputation. The most predictable variables are Hemoglobin at admission, the number of previous readmissions, and age. The model shared the number of previous readmissions as the most predictable variables as the model without imputation. However, imputation for missing data did not improve the prediction accuracy of the model. The number of previous readmissions is essential for predicting 30 days heart failure readmission across the different machine learning algorithms with and without imputation.

1.3 CONTRIBUTIONS

In this dissertation, I shed light on an overlooked yet important group of readmission patients. It is the first attempt to characterize day 1-30 all-cause readmission for heart failure patients to find subgroups for visits shorter than 30 days, with the aim of defining very early readmission. The future goal of my study is trying to minimize very early readmission. This group of patients can result in a disproportionate number of visits afterward that is associated with the increase in both the risk of mortality and the cost of treatment. (Kind, et al., 2008) While characterizing patients with the risk of all-cause or heart failure only readmission has been investigated a lot in the literature, to my knowledge, no work has been done to characterize d1-30 readmissions and find the parameters that could help predict very early readmission. My work is innovative because I hypothesize that patients within 30 days of readmission will have different patient characteristics and should not follow the same discharge protocol as others. My research showed the importance of prediction models in an unexplored area, highlighting an unexamined group of heart failure patients. Currently, little is known about what causes heart failure patients to come back so quickly. My research aims to predict very early readmission patients before discharge and improve on the discharge decision making. It is a step toward personalized healthcare to improve patient care eventually. The model can reduce rapid readmissions and thus lower the risk of mortality of HF patients and potentially reduce the treatment cost.

The second contribution of my research is building a visit table that can be used for different analyses for heart failure patients. The code used to build the visit table could be generalized to different diseases since I am using ICD code and DRG codes to define the disease. That means researchers can use it to answer similar questions for different diseases or different questions for a similar disease. More generally, I seek to understand which of our steps and methods for creating our visit table might generalize to other institutions. To the extent that they generalize, I could build a tool that would help researchers automate the process of building an appropriate visit table, and reduce the need for time-consuming chart review. This tool would produce a “cleaned” visit table, which could then be the nucleus for many analyses, including prediction tasks.

Chapter 2. PREDICTING READMISSION, A REVIEW OF RELATED WORK

The increasing availability of Electronic Health Record (EHR) data in one hand, and the increasing prevalence of heart failure, on the other hand, give researchers an excellent opportunity to help reduce its epidemic and associated costs. Heart Failure has been considered an epidemic with significant mortality, morbidity, and healthcare cost. (Roger 2014) It is considered a public health problem, with the prevalence of at least 26 million people worldwide and it is increasing. (Lund, Rich, & Hauptman, 2018) There has been much research on predicting readmission and trying to minimize patient readmission within 30 days of discharge. One reason for the focus on 30-day readmission is because the Affordable Care Act (ACA) third stage started to penalize hospitals with less than 30 days readmission rate higher than the proposed excess readmission ratio. ("Hospital Readmissions Reduction Program (HRRP)," 2018) The research about predicting heart failure readmission varied between assessing readmission risk models and trying to identify patient characteristics before discharge that contribute to the variation in hospital readmission rates. (O'Connor et al., 2016; Ross et al., 2008) In my research, I focus on studies that built a prediction model to improve discharge decision making. Predicting Heart Failure readmission require accurate data about patients and knowing the factors that directly influence the risk of readmission.

In this chapter, I discuss some of the research analyzing readmission for heart failure patients. There are many risk models that researchers have built for Heart Failure to predict readmission and/or death within 30, 60, 180 days and one year of discharge. There are two great systematic reviews of readmission prediction models for heart failure patients; Ross et al. (2008) and a recent study by O’Connor et al. (2016). Also, the Acute Decompensated Heart Failure National Registry (ADHERE) was designed to close the gap in knowledge and care by prospectively studying characteristics, management, and outcomes in a wide-ranging sample of patients hospitalized with Acute Decompensated Heart Failure (ADHF). [Fonarow 2003] Multiple studies used the data from this registry to create models that stratify the risk of readmission and/or death. Moreover, The Organized Program to Initiate Lifesaving Treatment in Hospitalized Patients with Heart Failure (OPTIMIZE-HF) registry was also created to bridge the gap between knowledge and medical care through the use of evidence-based research. Just like ADHERE, researchers used the OPTIMIZE-HF registry to build models that help them predict/stratify the risk of readmission and/or death.

However, shorter than 30 days readmissions ranges have received less research attention, which is the theme of my research. To my knowledge, there has been no prior research that has tried to define who is a rapid (or very early) readmission patient with heart failure disease and what patient characteristics could lead to very early readmission (shorter than 30 days). I hypothesize that very early re-admission are those readmissions that are more likely due to error, or readmissions, that if avoided, could have led to substantial cost savings and/or potentially improved health outcomes. In the following paragraphs, I will talk about the above four resources: The two systematic reviews (Ross et al. and O’Connor et al.) and the two registries (Adhere and OPTIMIZE) in detail. Then, I will discuss my systematic review trying to locate publications that analyzed shorter readmission ranges for heart failure patients.

2.1 PREVIOUS WORK

2.1.1 *Ross et al. 2008*

In a systematic review, Ross et al. (2008) included Ovid MEDLINE, PubMed, Scopus, Ovid PsycINFO, and all Evidence-Based Medicine Reviews on Ovid databases from the period of January 1, 1950, to November 19, 2007. In their search, they used the following Mesh Terms; patient readmission, risk, and heart failure, congestive. This yielded 941 articles. Then, 824 publications were excluded based on their inclusion and exclusion criteria (see table 1). From the remaining 117 publications, Ross et al. selected five papers that presented statistical models that either was derived or built to predict the risk of readmission for heart failure patients; Chin and Goldman 1997, Philbin and DiSalvo 1999, Krumholz 2000, Felker 2004, Yamokoski 2007. (see table 2.2) The other 112 studies that analyzed patient characteristics associated with readmission for Heart Failure patients did not derive a statistical model to predict the patient risk of readmission or compare hospital rates of readmission. None of the 112 studies presented models that predict the risk of readmission. Ross et al. concluded from analyzing the five papers that no consistent predictors appeared from their review. This could be because of the methodological heterogeneity among studies in the analytic approach, outcome examined and followed up periods, and handling death and transfer among patients. Moreover, the risk of readmission was high after HF hospitalization among the different studies in both short and long follow up periods, which emphasizes the importance of focusing on HF readmission. The result from Ross et al. gives me the idea to mimic the setup of other models (Adhere and OPTIMIZE) and use my dataset to see if it will yield similar predictors variables as the one these models got. Ross et al. compared results from different studies with the different model set up (such as variables, statistical method, and range of readmissions). I want to know if replicating the model set of (Adhere and Optimize) but with different dataset might yield a different result.

Inclusion	Exclusion
Studies from 1950 – November 19, 2007	Abstracts, Pediatric studies, Non-English language studies
Publications reported on readmission among HF patients as the primary or secondary outcome.	Publications without original data
	Studies with reported results from a case series or case report.
Studies that use data collected from a randomized Clinical trial that examines the effect of patient characteristics on readmission.	Studies from Experimental Studies, randomized clinical trials that focus on the effect of the intervention at readmission,

Table 2.1: Ross et al. (2008) Inclusion and Exclusion Criteria

Study	Study Type	Study Outcome	Readmission Period	Study significant Predictors	C-Statistic Score
Chin and Goldman, 1997	Prospective cohort	All-cause readmission or death	60 days	Number of HF hospitalizations within one-year, elevated BUN, lower systolic blood pressure, decreased hemoglobin, and a history of percutaneous coronary intervention (PCI)	Not Provided
Philbin and DiSalvo, 1999	Retrospective cohort	HF-specific readmission	One year	Patients treated at rural hospitals, patients discharged to skilled nursing facilities, and patients with echocardiograms	0.60

				or cardiac catheterization were less likely to be readmitted	
Krumholz et al. 2000	Retrospective cohort	All-cause readmission	180 days	Prior admission within one-year, prior heart failure, diabetes, and creatinine level >2.5 mg/dL at discharge	0.64
Felker et al., 2004	RCT cohort	All-cause readmission or death	60 days	Number of HF hospitalizations in the preceding 12 months, elevated BUN, lower systolic blood pressure, decreased hemoglobin, and a history of percutaneous coronary intervention (PCI)	0.69
Yamokoski et al. 2007	RCT cohort	All-cause readmission and death	Six months	serum BUN level and high-dose diuretics at discharge	0.60

Table 2.2: List of the five prediction models Ross et al. selected from the literature and their characteristics

The five papers that Ross et al. (2008) have selected [see table 2.2] analyzed either all-cause readmission or death or HF readmission alone. The readmission ranges varied from 60 days to one year. These five models performed poorly, and some of them have not reported their prediction performance. Moreover, the fact that there was not a consistent predictor among the studies could be because of the different analytic approach, outcome examined and followed up periods, and the way death and transfer among patients were handled among the different studies.

2.1.2 *O'Connor et al. 2016*

A more recent systematic review by O'Connor et al. (2016) aimed to identify heart failure patient characteristics measured before discharge that contribute to the variation in hospital readmission rates. Their literature search focused on studies that investigated readmission within 7 – 180 days. O'Connor et al. have searched CINAHL, PubMed, and Cochrane databases for the period of January 1992 and September 2014. In their search they used the following Mesh Terms; Heart Failure, Patient Readmission, and Hospitalization with different synonymous. This yielded 950 abstracts. Then, 805 abstracts were excluded based on their inclusion and exclusion criteria (see table 3). From the remaining 145 publications, O'Connor et al. have conducted a full-text review and further eliminated 111 articles. O'Connor selected 34 studies that met their eligibility criteria to conduct their review on. Also, O'Connor et al. work yielded a similar result as Ross et al. with not finding a single patient characteristic considered as the key factor among the different studies. Again, the reason for not finding a single key factor characteristic could be similar to Ross et al. with different analyses and readmission ranges. O'Connor et al. included only one paper that analyzed shorter than 30 days readmission.

This underlines a challenge in developing a successful prediction model to reduce readmission. The differences in model setup could be one reason behind not finding a single patient characteristic across the studies. Therefore, it is crucial to build a generalized heart failure readmission prediction model and be able to find patient factors associated with heart failure readmission. This could be done by mimicking the setup of current models and check if that will yield different outcomes.

Inclusion	Exclusion
Studies between January 1992 – September 2014	Non-English language studies
Publications on HF patients treated outside the US.	Abstracts with incomplete information
Reported outcomes within six months	Studies examining only post-acute mortality or a combined outcome of readmission and mortality.
Identified statistically significant patient risk factors for readmission.	

Table 2.3: O'Connor et al. (2016) Inclusion and Exclusion Criteria

2.1.3 *The Acute Decompensated Heart Failure National Registry (ADHERE):*

ADHERE registry was created to improve the care and better define acutely decompensated heart failure (ADHF) patients, built as a national, multicenter study of patient characteristics, the pattern of care, and outcomes of patients admitted with ADHF. There were 65,150 patients enrolled in the registry from 263 hospitals for the period from October 2001 to July 2003. The main eligibility criteria for the registry is if the patient was admitted to the acute care hospital with a previous discharge of heart failure diagnosis. (Fonarow, 2003) After searching PubMed, I found about 42 publications used the Adhere registry for their research questions. The publications research questions vary from, predicting in-hospital mortality, measuring the quality of care, analyzing patients' characteristics, and hospital readmission.

Fonarow et al. in 2012 built a bedside tool for risk stratification for patients hospitalized with ADHF. I choose this paper, in part, because they provided a detailed explanation of the setup of the prediction model. It is essential to have a detailed model setup in order to mimic their work using my dataset. Fonarow et al. used the data from the ADHERE registry as their cohort with variables predicting materiality in ADHF as the primary outcome to measure. Their analysis was based on 65,275 patient records from the registry. For their analysis, they used the classification and regression tree (CART) analysis to find the best predictors of in-hospital mortality and to develop the risk stratification model. The CART is a statistical method based on recursive partitioning analysis. [Brelman 1984, Yohannes 2004] It can handle highly skewed numerical data and categorical predictors (ordinal or nonordinal) since it does not require parametric assumptions. [Fonarow 2012] Fonarow et al. analyzed a variety of variables in their study; demographics, type of insurance, heart-failure history, medical history, laboratory values, and initial vital signs. From the 39 variables that they included in their model, blood urea nitrogen (BUN) at admission is the best discriminator between survivors and non-survivors (at 43 mg/dl or higher). The next best predictor of in-hospital mortality is systolic blood pressure (SBP) where the decision tree splits at a level of less than 115 mg Hg. Then, the third best predictors that discriminate survivors from non-survivors is Serum Creatinine at a level of 2,75 mg/dl or higher. Analyzing these predictors is vital for chapter four when I build my model to see if using these 39 variables that Fonarow et al. has used will result in the same predictors and their levels.

2.1.4 *The Organized Program to Initiate Lifesaving Treatment in Hospitalized Patients with Heart Failure (OPTIMIZE-HF) registry:*

OPTIMIZE-HF registry was created to increase the use of evidence-based therapies given the availability of patient data to reduce mortality and readmission rates for heart failure patients. It aims to create a better understanding of the barriers to using beta blockers and An angiotensin-converting-enzyme inhibitor (ACEIs). Moreover, the OPTIMIZE-HF program is aimed at guideline-recommended therapies for HF patients before discharge. (Fonarow et al., 2004) The data from the registry was collected from 259 U.S hospitals. The Inclusion criteria for patients were eligible are; hospitalized were the primary cause of admission is worsening HF or significant symptoms of HF at admission, systolic dysfunction (LEVF <40%), and patients with ICD-9 heart failure code of 402.01, 402.11, 402.91, 404.01, 404.11, 404.91, 428.0, 428.1, 428.9.

The data was collected between March 2003 and December 2004 and 48,612 patients were enrolled.

O'Connor et al. 2008 conducted an analysis using an OPTIMIZE-HF registry that aims to locate the predictors of mortality after discharge in HF patients. In their analysis, they collected follow-up data at 60 and 90 days postdischarge from prespecified 10% sample. From the 259 hospitals, 91 hospitals participated in the follow-up and data was collected from 5791 patients. From their cohort, the average number of days of follow-up after discharge is 72.7 +- 21.5. The 60 to 90 days postdischarge mortality was 8.6%, and 29.6% were rehospitalized. They have included 45 potential variables in their logistic regression model. 13 of the 19 predictable variables that were analyzed found to be predictive of mortality. These factors are age, serum creatinine, reactive airway disease, liver disease, lower systolic blood pressure at admission and discharge, lower serum sodium, lower admission weight, lower extremity edema, and depression. The best predictors for both mortality or rehospitalization were serum creatinine at admission, systolic blood pressure, hemoglobin at admission, use of angiotensin-converting enzyme inhibitor or angiotensin receptor blocker at discharge, and pulmonary disease. The C-Statistic score for their model is 0.723 for postdischarge mortality or rehospitalization. The C-Statistic score for mortality within 60 days is 0.72. Moreover, the study concluded that the use of beta blockers and statins at discharge was associated with reducing mortality.

A similar study by Abraham et al. 2008 was conducted to develop a predictive model for in-hospital mortality to heart failure patients and locate the main predictor variables. From the OPTIMIZE-HF registry 48,612 heart failure patients, there were 1,834 (3.8%) in-hospital mortality incidents. The study removed patients with missing/incomplete variables. This removal resulted in 37,548 patients and 1217 deaths. For their analysis, they used a point scoring system calculated from the seven most important predictors from a multivariable logistic regression analysis. These predictors are serum creatinine (SCr) at admission, systolic blood pressure (SBP) at admission, age, Sodium, prior cerebrovascular events, heart rate, and beta blocker at admission. The C-Statistic score for their model was 0.77. Moreover, the study found that a patient who took beta-blockers or angiotensin-converting enzyme inhibitor at admission will have a lower risk of in-hospital mortality.

Moreover, the work that has been done on both ADHERE and OPTIMIZE-HF registries was mainly on predicting in-hospital mortality, not readmission. Also, they mostly looked 30 days to one-year mortality and/or readmission. Using OPTIMIZE-HF, Kociol et al. 2011 looked into predicting one-year mortality or/and readmission among older patient (>=65 years) hospitalized with heart failure to see which measure of B-type natriuretic peptide (BNP) is the most important predictor- admission, discharge, or the difference between admission and discharge. The study yielded that BNP at discharge is the most informative for one-year mortality or readmission. (Kociol et al., 2011)

2.2 DISCUSSION

From the previous work above, it is clear that there has been plenty of work done on risk stratification of death or readmission for heart failure patient. The above two systematic reviews and the work done utilizing the two registries have looked into ranges between 30 days to one year. To my knowledge, there is very little research that looks into readmissions that are shorter than 30 days. Also, it is not clear what duration should be considered ‘short’ readmission what characteristics are associated with that. Gabayan et al. defined all cause rapid readmissions to the emergency department to be within seven days. (Gabayan et al., 2013) Moreover, Eastwood et al. have looked up the determinant factors that would increase the likelihood of early readmission after HF hospitalization, focusing mainly on 7 and 30 days of readmission. The paper claimed that readmission between 0-6 days after discharge is considered more avoidable than admission within 20-27 days. Also, similar to Gayan et al. who looked into rapid emergency room readmission, they selected seven days because of an analysis that was done by Clarke in 1990 where he assigned 0-6 days avoidable readmission in general not specific for heart failure. Clarke analysis was not specific for heart failure patients but all types of readmission including general surgical inpatients. (Eastwood et al., 2014) (Clarke, 1990)

In my study, I am seeking to locate a set of modifiable characteristics that could affect the risk of readmission rather than one single attribute. Eastwood et al. (2014) was the only study that looked at less than 30 days readmission when it looked at 7-day readmission. They have looked into determining the factors that could increase the risk of readmission within seven days and 30 days. (Eastwood et al., 2014) In that study, they found that 5.6% of their 18,560 HF patients were readmitted within seven days. The study concludes that these seven days of readmission is associated with the history of kidney disease. The study has neither looked into the cardiac severity such as ejection fraction nor the type of medication used inside the hospital. I believe that measuring the severity of the disease and factoring the medication intake inside the hospital is important for predicting readmission.

2.3 AN UPDATED REVIEW OF THE LITERATURE

I conducted a review and summarized the literature on the area of heart failure readmission prediction by combining the knowledge gained from reviewing Ross and O’Connor systematic reviews and the work that was done using ADHERE and OPTIMIZE-HF registries. Unfortunately, short readmission was rarely investigated. I started by defining my patient group and type of models. In my research, I focused on models that either predict readmission alone or readmission and death, excluding models that use death as their only output measure. I focused my search on including the recent studies and not to duplicate the Ross and O’Connor work. For that, I have searched PubMed for the period of October 2014 and December 2017. In my search, I used the same Mesh Terms that O’Connor used in their systematic review: ([“Heart Failure” OR “Cardiac Failure” OR “Congestive Heart Failure” OR “Heart Decompensation” OR “Heart Failure, Congestive” OR “Heart Failure, Left-Sided” OR “Heart Failure, Right Sided” OR “Left-Sided Heart Failure” OR “Myocardial Failure” OR “Right-Sided Heart Failure”] AND [“Patient Readmission” OR “Hospital Readmission” OR “Hospital Readmissions” OR “Readmission, Hospital” OR “Readmissions, Hospital”] AND [“Hospitalization”]). Then, I added these Mesh Terms; [“Short Readmissions” OR “Rapid Readmissions” OR “Avoidable Readmissions” OR “Unplanned Readmissions” OR “Early Readmissions”]. This yielded very few abstracts that are related to my research. Most abstracts refer to early, short, or avoidable readmission to be within 30 days. The following table list my inclusion and exclusion criteria for publications.

Inclusion Criteria	Exclusion Criteria
Studies between October 2014 – December 2017	Non-English language studies
Heart Failure as the Primary or Secondary Reason For Readmission.	Case Report or Case report studies
Identified statistically significant patient risk factors for readmission.	Studies that predict death only.
Identify patient risk factor	No quantitative results (such as no C-Statistic Score)
	Therapy promoting studies.

Table 2.4: My Inclusion and Exclusion Criteria

Study	Study Type	Study Outcome	Readmission Period	Study significant Predictors	C-Statistic Score
Minana et al. 2017	Prospective Cohort Design	All-cause readmission after an earlier acute heart failure	10 and 15 days	Length of Stay (LOS) with 4 days or less Longer than 7 days LOS	0.72 and 0.70 for all-cause readmission
Dharmarajan et al. 2013	Retrospective cohort	All-cause readmission after hospitalization for HF, Myocardial Infraction, or Pneumonia and the relation of patient demographics Characteristics to readmission period	0-3 days 0-7 days 0-15 days	Neither readmission diagnoses nor timing substantively varied by age, sex, or race	N/A
Amarasingham et al. 2010	Retrospective cohort	HF-specific readmission or death	30 days	Markers of social instability and lower socioeconomic status, demographics, health behavior	0.86 (death) 0.72 (readmission)
Eastwood et al. 2014	Retrospective cohort	All-cause Heart Failure	7 days 30 days	Age, history of kidney disease,	0.73

				Discharge Disposition category	
Betihaves et al. 2015	A prospective, multicenter randomized controlled trial	Unplanned cardiovascular readmission after a hospitalization for CHF	Within 28 days After 28 days	Age, Living Alone, Sedentary lifestyle, the presence of multiple comorbid conditions	0.80

Table 2.5: List of the five prediction models I selected from the literature and their characteristics

There has been extensive work done in the literature trying to predict readmission for heart failure patients. Researchers have studied different temporal definitions of "readmission," from 7 days to one year. However, very few papers analyzed the patient characteristics with readmissions of less than 30 days. Researchers use the terms short, early, rapid, and avoidable readmissions to describe readmissions that happen within 30 days. Some papers talked about shorter than 30 days of readmission frequencies. Dharmarajan et al. investigated the frequency of readmission following heart failure, Acute Myocardial Infarction, or Pneumonia hospitalization for the periods; 0-3 days, 0-7, and 0-15 days. From the 329,308 30-Days readmissions after a HF hospitalization, 13.4% 0-3 days, 31.7% 0-7 days, and 61.0% 0-15 days readmission. The paper concluded that patient demographics; age, sex, and race did not affect the comorbidity-adjusted hazard ratio.

Moreover, Amarasngham et al. built a real-time electronic predictive model to identify hospitalized heart failure patients at high risk for readmission or death. The model C-Statistic score for predicting death was high at 0.86 and within range of 0.72 in predicting readmission within 30-days. The study found that incorporating the social deterministic factors in the model increased its accuracy. Comparing to other models, their model performed better than the ADHERE model which yielded a C-Statistic score of 0.56. In their analysis, they classified the risk category into; very low, low, intermediate, high, and very high. They found that patients with higher risk category were readmitted earlier within the 30-days postdischarge range. (see figure 2.1).

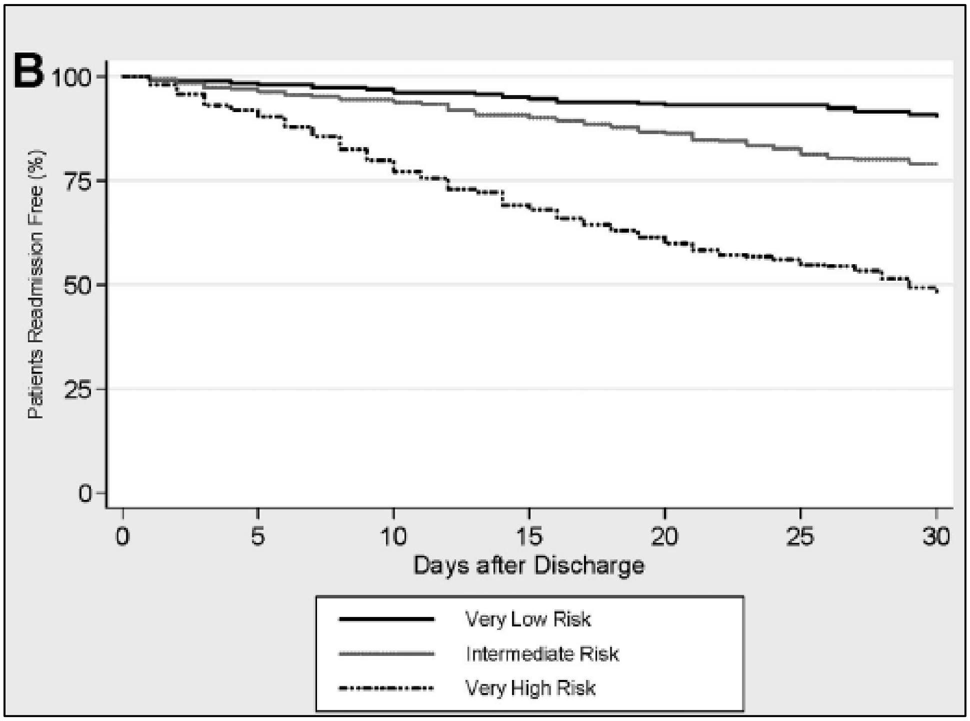


Figure 2.1: Percent of patients' readmission free over the 30-day post discharge interval stratified by very high, intermediate, and very low risk. (Amarasingham et al. 2010)

Minana et al. 2017 is a recent study that looked into the relationship between the length of stay and the risk of very early readmission for heart failure patients. They grouped the length of stay (LOS) into; ≤ 4 days, 5-7 days, 8-10 days, and >10 days. They investigated the association between these groups and 10, 15, and 30 days of readmission. The study did not define what is considered short readmission and why but concluded that short hospitalization is inappropriate even if it is the first admission. Minana et al. showed that patients with LOS ≤ 4 days had increased the risk of early readmission. The study explained their outcome to be the result of; 1) clinician couldn't easily recognize the complications or they might not be present during the early days of admission, 2) the shorter hospitalization didn't allow clinician to perform thoroughly diagnostic exam, and 3) the shorter hospitalization didn't allow clinician to carry out an optimal pharmacological titration. (Miñana et al., 2017) Also, longer length of stay (>7 days) is associated with an increased risk of very early readmission. The C-Statistic score for their model were 0.72 and 0.70 for all-cause readmission for 10 and 15 days post discharge.

2.4 CONCLUSION

To conclude, there is very little prior research that investigates early readmission (less than 30 days), nor defines ideas such as rapid, short, early, or unplanned readmissions of heart failure patients. Also, little is known about patient characteristics associated with early readmission. Eastwood et al. have looked up the determinant factors that would increase the likelihood of early readmission, focusing mainly on 7 and 30 days of readmission. (Eastwood et al., 2014) Moreover, Minana et al. looked into one single factor and its association for early readmission that is the length of stay. The study picked 10, 15, and 30 days of readmission as follow up end pints with no justification of doing so. Also, Dharmarajan et al. investigated the association of patient demographics to early readmission. Same as Minana et al. study, Dharmarajan et al. chose three readmission ranges; 0-3 days, 0-7 days, and 0-15 days with no justification of choosing so.

An important outcome of my research is to define the range of days for shorter readmission and understand the patient characteristics that help predict it rather than picking one variable and see its effect on early readmission. Also, I seek to find the differences in patient characteristics within 30 days if it exists. For example, patient characteristics for readmission within seven days could be different from readmission within 25 days. This will help discover if the literature definition of early readmission within 30 days is scientific and due to clinical factors or not? I hypothesize that there exist a different group of patients within 30 days of readmission (shorter than 30 days) and thus should be treated differently when we analyze early or avoidable readmissions.

Currently, little is known about what causes heart failure patients to come back so quickly. My proposed work is to predict rapid readmission patients before discharge and improve on the discharge decision making. It is a step toward personalized healthcare to improve patient care eventually. The model can reduce rapid readmissions and thus lower the risk of mortality of HF patients and potentially reduce the treatment cost.

Chapter 3. COUNTING READMISSIONS: SURPRISINGLY DIFFICULT

The first research question that needs to be addressed is what counts as readmission when creating a visit table for cardiac patients. This is not a trivial problem for researchers as noise does exist in the data due to hospital policies, billing procedures or variation in care practices. For researchers to build an accurate visit table, they must clean and validate the data through different resources such as the identified EHR data, discharge summary report, and manual chart review for some patients.

The availability and accessibility of Electronic Health Record (EHR) data create an opportunity for researchers to revolutionize healthcare. The recognition of the importance of secondary use of EHR data has led to the development of research-ready integrated data repositories (IDRs) from EHR data. Analyzing this data can help researchers connect the dots and can lead to important clinical findings through predictive analytics methods. Unfortunately, poor data quality is a problem that affects the accuracy of such findings. An example of a data quality problem is poor information about the specifics of admission, discharge, and readmission.

I would like to leverage the availability of data to understand readmission patterns for heart failure patients. To understand readmission patterns, I need an accurate visit table with patient admission and discharge data which can then be used to query for individuals' readmission history, or for institution-wide patterns. Unfortunately, I found that building such a visit table is surprisingly difficult. If noise or inaccuracies exist in this table, that will affect the accuracy of any predictive analytics that might be built from this information. This is an important question to answer before tackling my second research question (as described in chapter 4), which aims to find the differences in characteristics for patients who are readmitted within 30 days. Below, I describe some considerations, challenges, and methods for building such a visit table.

In the process of seeking an answer for my first research question, I started by analyzing the "research-ready" integrated data repository (IDR) that currently exists at the University of Washington. These are copies of the EHR data that is De-identified and build to facilitate the secondary use of data for researchers. (see figure 1) Then, I analyzed the identifiable data and compared the results to see how accurate my visit table is and if the de-identifying process could affect the accuracy.

3.1 METHOD

In this section, I started with a preliminary analysis using the De-identified Clinical Data Repository (DCDR) which contains a subset of Caradigm data. Caradigm is an aggregation of data stored across a broad collection of UW medicine health systems, including ORCA for inpatient data, EPIC for outpatient data and some other systems. The work on DCDR data allowed me to understand the field and the data that exist in the Electronic Health Record data. Then, I wanted to see if the data quality issues that affected my analysis with DCDR data is due to the de-identification process. For that, I conducted a similar analysis but with the whole Caradigm data to see if the data quality issues that affect the accuracy of the visit data is an actual error that exists in the EHR data.

3.1.1 *Preliminary Analysis Using De-Identified Data and Identified-Data*

I started my analysis using the De-Identified Clinical Data Repository (DCDR) which contains a subset of Caradigm data. Caradigm is an aggregation of data stored across a broad collection of

UW medicine health systems, including ORCA for inpatient data, EPIC for outpatient data and some other systems. DCDR is an anonymized version of Caradigm. Figure 3.1 shows the flow of clinical data for use by researchers at the University of Washington Medicine. The data collected in the DCDR goes back to 1994 and up to 2017. The DCDR is a cohort identification/feasibility estimation tool. (De-Identified Clinical Data Repository (DCDR), n.d.) I interface to DCDR through a secure web-based query tool powered by i2b2 [Figure 3.2] (i2b2 Query and Analysis Tool, 2016).

I started querying for heart failure using ICD-10 codes. The class "Heart Failure" has a large number of subclasses such as diastolic, combined systolic, systolic heart failure and more. In this stage of my analysis, I consulted a cardiologist who highlighted specific Diagnosis Related Groups (DRGs) associated with heart failure. These DRGs translate into the following ICD 10 codes; I09.81, I11.0, I13.0, I13.2, I50.1, I50.20, I50.21, I50.22, I50.23, I50.30, I50.31, I50.32, I50.33, I50.40, I50.41, I50.42, I50.43, I50.9, R57.0, R57.9. (See Table 3.2 in section 3.2.2) (Heart failure I50, n.d.) I refined the data by adding discharge disposition concepts. There are different dispositions associated with the discharge (see Table 3.1) such as Hospice, Died/Expired, Left against Medical Advice, Nursing Facility, and Psychiatric facility. I discarded data that would be inappropriate to include for analysis of readmission patterns. For example, if a patient leaves against medical advice, and later is readmitted, that readmission should not count for analysis, if the aim is to build a system to improve physician discharge decisions. Also, other discharge disposition concepts were discarded, such as ones that are not going to have readmission like "Died/Expired," or are not considered acute medical service readmission like "Psychiatric facility" and "Rehab facility."

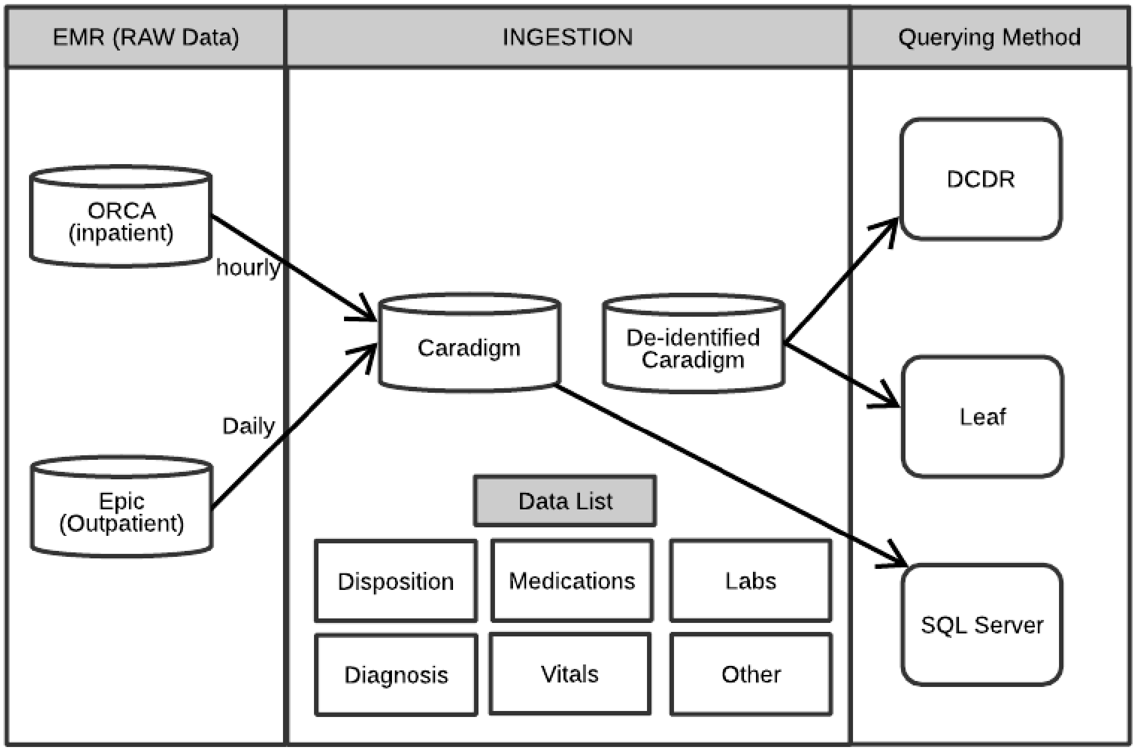


Figure 3.1: The flow of clinical data for use by researchers at the University of Washington Medicine

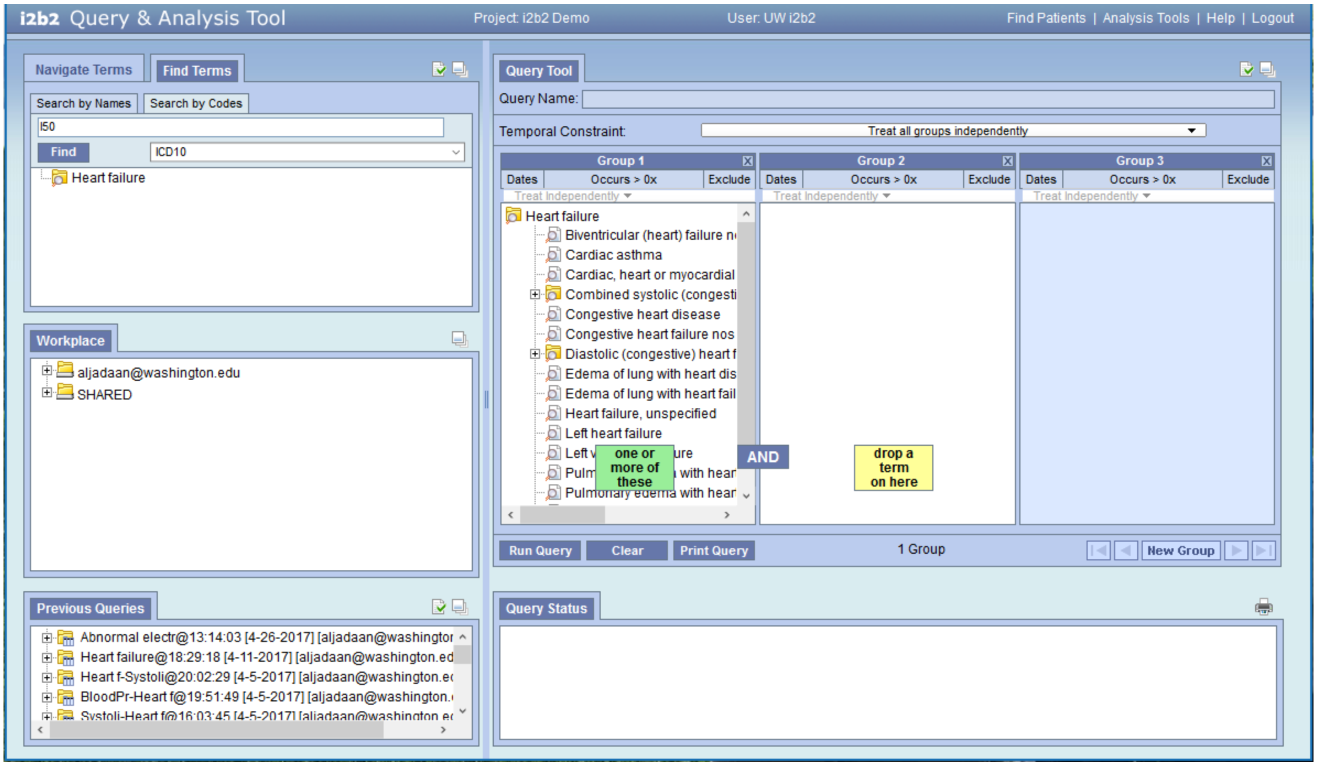

i2b2 Query & Analysis Tool
Project: i2b2 Demo
User: UW i2b2
Find Patients | Analysis Tools | Help | Logout
Navigate Terms
Find Terms
Search by Names
Search by Codes
I50
Find
ICD10
Heart failure
Workplace
aljadaan@washington.edu
SHARED
Previous Queries
Abnormal electr@13:14:03 [4-26-2017] [aljadaan@washington
Heart failure@18:29:18 [4-11-2017] [aljadaan@washington.ed
Heart f-Systoli@20:02:29 [4-5-2017] [aljadaan@washington.e
BloodPr-Heart f@19:51:49 [4-5-2017] [aljadaan@washington.
Systoli-Heart f@16:03:45 [4-5-2017] [aljadaan@washington.e
Query Tool
Query Name:
Temporal Constraint:
Treat all groups independently
Group 1
Group 2
Group 3
Dates
Occurs > 0x
Exclude
Treat Independently
Heart failure
Biventricular (heart) failure n
Cardiac asthma
Cardiac, heart or myocardial
Combined systolic (congesti
Congestive heart disease
Congestive heart failure nos
Diastolic (congestive) heart f
Edema of lung with heart dis
Edema of lung with heart fail
Heart failure, unspecified
Left heart failure
Left v
ure
Pulm
with hear
Pulmonary edema with hear
one or more of these
AND
drop a term on here
Run Query
Clear
Print Query
1 Group
New Group
Query Status

Figure 3.2: The i2b2 Query & Analysis Interface

No.	Discharge Disposition	No.	Discharge Disposition
1	Home/Self Care	10	Dischrg/Tr: Disch/Trans Fed Hospital
2	Skilled Nursing Facility	11	Hospice
3	Against Medical Advice	12	Disch/Trans/Planned Readmission to Long Term Care Hospital
4	Expired /Dead	13	OTH INST: Other Institution - Not Defined Elsewhere
5	Disch/Trans to a Distinct Psych Unit	14	STILL A PATI: Still a Patient
6	Disch/Trans to Court/Law Enforcement	15	CA CTR/CHLD: Designated Cancer Center or Children's Hospital
7	Disch/Trans to a Distinct Rehab Unit/Hospital	16	DECEASED - O: Deceased - Organ Donor

8	HOME HLTH: Home Health Care	17	Disch/Trans/Planned Readmission to ICF-Intermediate Care Facility
9	TRANSFER TO: Transfer to Hospital	18	Disch/Trans/Planned Readmission to Other Institution-not defined elsewhere

Table 3.1: Types of Discharge Dispositions

In this preliminary analysis, I have selected 2,000 patients from heart failure patient group based on the ICD code definition. EHR Data suffers just like any other type of data in the fact that it is messy and incomplete and requires a lot of data cleaning. For that, I created a script in R to set up the data and conduct data cleaning when needed. Also, the R script calculates certain variables that I believe are vital to measuring when building the visit table. First, I calculated the days until readmission, measured by subtracting the time stamp of the second readmission from the time stamp of first discharge. Then, I calculated the total number of readmissions: simply counting the total number of readmissions per patient. Also, I calculated the length of stay per visit: the date of discharge minus the date of admission. Since DCDR includes inpatient and outpatient data, and my study focuses only on inpatients, I deleted visits where the length of stay is 0 days. This is because the DCDR does not have a simple flag indicating inpatient versus outpatient data. After cleaning the data, the total number of patients is 1807.

Before I start making assumptions about patient visit patterns, I needed to test the accuracy of the visit table that DCDR provides through the admission and discharge timestamp and the discharge disposition (ex. home, hospice, nursing facility). I wanted to check the quality of the visit data on a small set of patients to allow me to do manual chart reviews easily. The preliminary results that I found are that there are 45 patients from the 1807 patients (%2.5) who have readmission within the same calendar day of their previous discharge. For that, I started looking at the readmission that happens within the same calendar day as the previous discharge. After consulting with a cardiologist, he concluded that the percentage of these same day readmission patients is abnormal. The distribution of the number of hours of the same day readmission shows that huge bulk of the cases happens within the first 7 hours. [Figure 3.3] The interesting part is for cases of readmissions that occur within the first 1-3 hours. I wanted to make sure if these are actual readmission cases as marked in the EHR or they are transfer cases that were mislabeled as discharge to home.

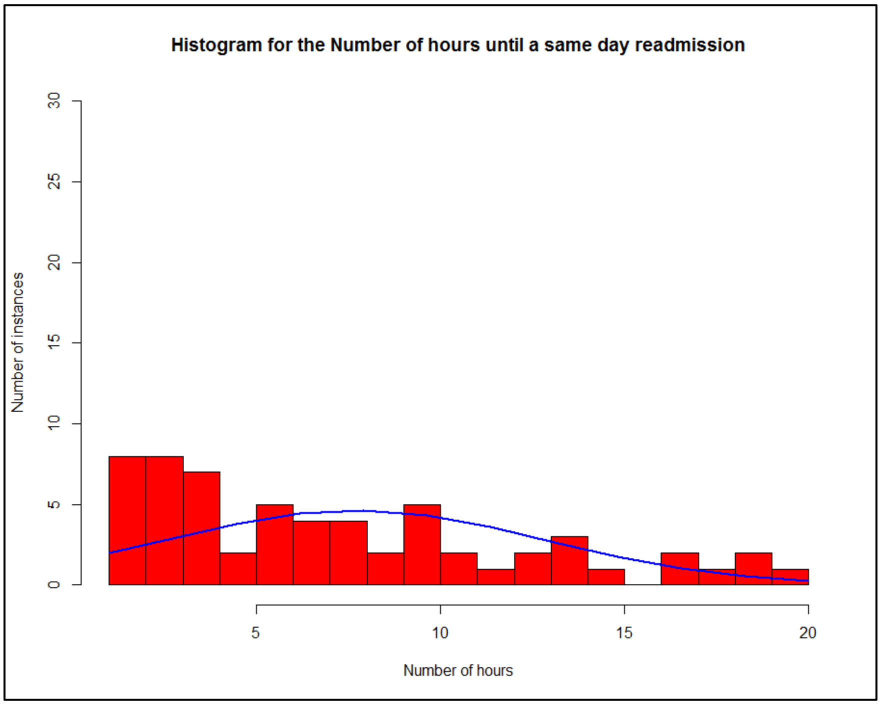


Figure 3.3: Histogram of hours until the same-day readmission using DCDR dataset

The data noise I was able to detect from the preliminary work with DCDR dataset encouraged me to apply the same methods when I queried Caradigm data. I wanted to see if the data quality issues detected from DCDR is due to the de-identification process that affected the accuracy of the visit data or an actual error that exist in the EHR data. This preliminary work helped me locate data quality issues that exist in the EHR data that need to be cleaned before conducting the analysis such as null values, the unorganized order of visits, and duplicated values, acute vs. non-acute readmissions, and discharge dispositions that do not affect the physician discharge decisions.

I started querying identified EHR data from Caradigm. Absent from the friendly query interface that DCDR provides, I had to query the whole patient's visit table from Caradigm and the dataset for Diagnosis Related Group (DRG) using SQL server. I got two data files; one for the whole visit table and the other is for the DRGs codes associated with visits IDs. Then, I created an R script to help me with my cleaning and analysis. First, I only kept DRG codes that are heart failure related: "MS291", "MS292", "MS293", "AP127", and "APR194". Then, I joined the two data files, where I selected patients in the large visit table that have these DRG codes. The data is not clean to conduct my analysis, so I did some data cleaning, which includes deleting duplicated and Null data, ordering the visit date per patient. After that, I deleted visits that happen before the initial diagnosis of heart failure, keeping only visits that happened after the initial heart failure diagnosis. So, the new visit table for each patient started with the heart failure related admission and followed by any cause readmission. The goal of creating a visit table is to analyze day 1-30 readmission patients and see if their early discharged could have led to very early readmission of any cause. After joining the two data files and cleaning the data, the new visit table now has 1832 heart failure patients with 5600 visits.

Then, I wanted to apply the same technique that I did earlier on the DCDR dataset to ensure the consistency in my process. Just like my process in the DCDR, I used the same discharge deposition concepts. Checking patients with readmission that happened on the same calendar day of the previous discharge. This resulted in 48 visits/observations for 47 patients (~2.6%). [Figure

3.4]. Some of these visits have Account Status that says “Discharged” and DemoDischargeDispositionDescr: “HOME: Home/Self Care,” yet the admit source description says, “Transfer w/in Hospital Resulting in Separate Claim to Payer.” Another example with some patients who have about 1-hour difference between the previous discharged and the next readmission. The “Unit” from the previous discharge is “Operating Room,” but the DemoDischargeDispositionDescr says “HOME: Home/Self Care.” It does not say transfer. Some of these findings have the potential to affect a reasonable proportion of apparent discharges, which would be reclassified as transfers.

The apparent differences between the two plots in figure 3 and 4 despite they are coming from the same data source raises a data quality flag. The average number of hours for same day readmission in the DCDR dataset is 7.5 hours whereas it is 3.5 hours from SQL_Server querying from CARADIGM. There should not have such a big difference in the shape and the average since we are querying the same UW Medicine dataset. There could be many reasons behind such differences including but not limited to the time-shift algorithm used to de-identify the patient. In reviewing the DCDR dataset, I found some instances where patients got discharged to home from the hospital between 10 p.m. - midnight and got readmitted the early morning of the next day between 1 a.m. – 3 a.m. In the healthcare norm, this rarely happens and could have happened due to the de-identify process used to de-identify the data or some human error. This error can have a huge effect on the research since these patients might have same day readmission, but with their current timestamp, they are not. This pattern of discharge and readmission did not appear as frequent on Caradigm data as it is in DCDR.

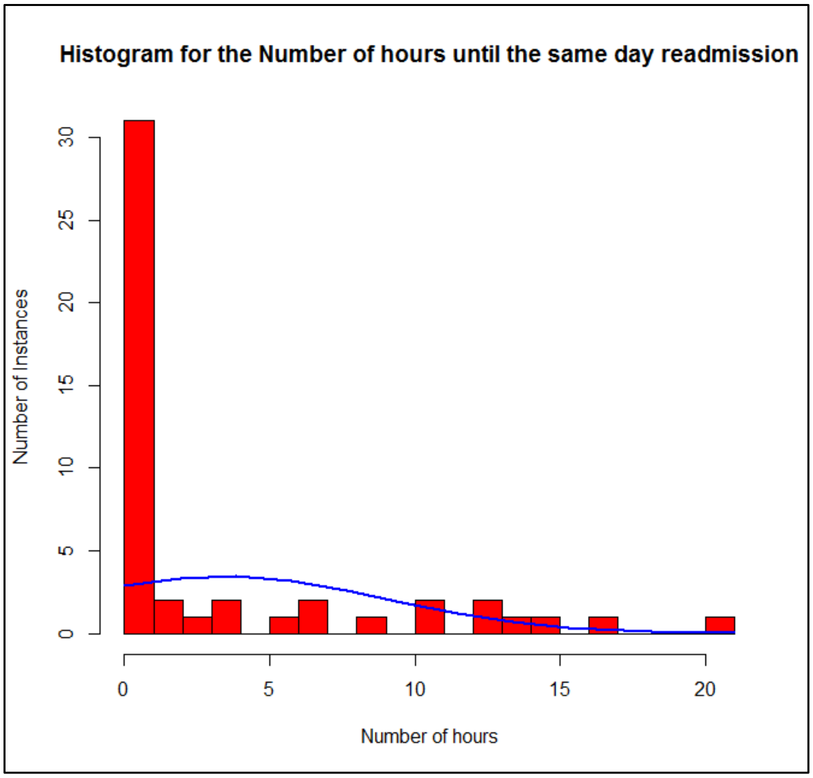


Figure 3.4: Histogram of hours until the same-day readmission using dataset queried directly from Caradigm

Another discrepancy exists due to the de-identification process when I try to locate 30 days readmission patients between the two datasets. In the DCDR dataset, there were 764 patients from the 1807 patient (~42%) who have readmission within 30 days with an average of 12 days. In the other dataset that is queried from Caradigm as well, I got 456 patients from 1827 patients (~25%) who have been readmitted within 30 days with an average of 13 days. These issues in the data and the discrepancy between the two datasets makes data quality assessment an important step to ensure that I have elicited an accurate visit before I carry on to my analysis and building my model. Giving the low quality of the DCDR dataset, I have decided to continue the remainder of my analysis using the identified dataset from Caradigm.

3.1.2 *Analysis Using Identified Data (Caradigm)*

The preliminary work I described in the previous section (3.2.1) gave me a better understanding of how to set up the data. I started by defining who is a heart failure patient and determined which ICD, CPT, DRG codes (Billing discharge code or physician-entered diagnosis code) and Laboratories scores associated with defining the heart failure patient. There exist different definitions of heart failure in the literature in the inclusion and exclusion of ICD, CPT, and DRG codes. The literature commonly uses the ICD code 428.0 that refers to heart failure, but they differ in including other ICD codes such as 786.5 Chest pain, 440 Atherosclerosis, and others. (Goff Jr, Pandey, Chan, Ortiz, & Nichaman, 2000) Also, some research studies defined the laboratories scores cutoff differently when defining their heart failure cohort. For example, a patient preserved ejection fraction (HFpEF), and reduced ejection fraction (HFrEF) score to be considered heart failure patient varies in the literature from <50 to <40. (Rutten, Clark, & Hoes, 2016)

Giving the different definitions literature defined their heart failure cohort, I have defined the heart failure cohort using a precise definition of Heart Failure patient that was provided by a cardiologist; Dr. Todd Dardas. The definition consists of ICD and DRG codes. Then, I translated the DRG to ICD 10 codes. The result is 20 ICD 10 codes that accurately define our heart failure population. (Table 3.2) In addition to the use of ICD and DRG codes, I have included some lab scores to the heart failure definition. This includes objective echo finding of Heart Failure either (abnormal diastolic function by echo OR Ejection Fraction <45% OR NT-proB-type Natriuretic Peptide Blood test (BNP) >200.

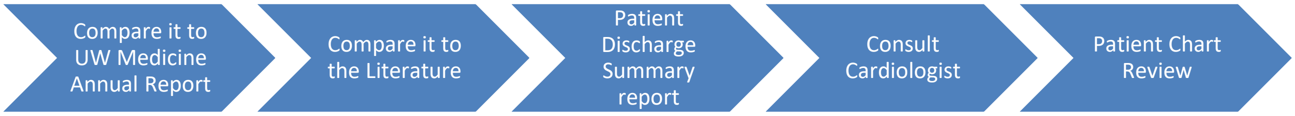
ICD 10 Code	Description	ICD 10 Code	Description
I09.81	Rheumatic HF	I50.31	Acute diastolic (congestive) heart failure
I11.0	Hypertensive heart disease with HF	I50.32	Chronic diastolic (congestive) heart failure
I13.0	Hypertensive heart and chronic kidney disease stage 1-4	I50.33	Acute on chronic diastolic (congestive) heart failure
I13.2	Hypertensive heart and chronic kidney disease stage 5	I50.40	Unspecified combined systolic (congestive) and diastolic (congestive) heart failure
I50.1	Left ventricular failure	I50.41	Acute combined systolic (congestive) and diastolic (congestive) heart failure
I50.20	Unspecified systolic (congestive) heart failure	I50.42	Chronic combined systolic (congestive) and diastolic (congestive) heart failure
I50.21	Acute systolic (congestive) heart failure	I50.43	Acute on chronic combined systolic (congestive) and diastolic (congestive) heart failure
I50.22	Chronic systolic (congestive) heart failure	I50.9	Heart failure, unspecified
I50.23	Acute on chronic systolic (congestive) heart failure	R57.0	Cardiogenic shock
I50.30	Unspecified diastolic (congestive) heart failure	R57.9	Shock, unspecified

Table 3.2: ICD codes used for Heart Failure Definition

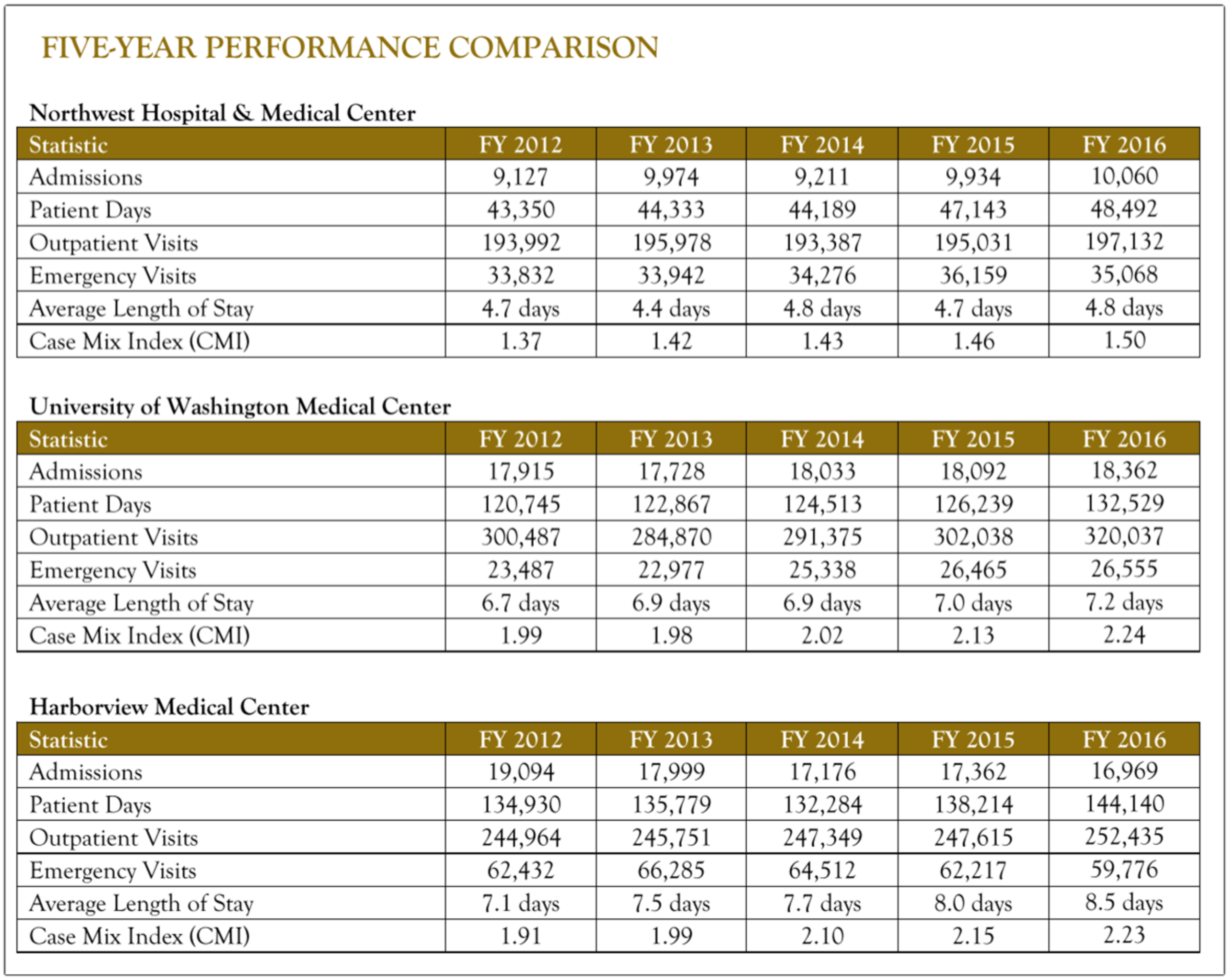
I started querying patients with the associated ICD codes using SQL_Server. In consultation with cardiologists and via manual chart review, we developed a set of rules to help “clean” the visit table. These rules successively discard data that would be inappropriate to include for analysis of readmission patterns. For example, if a patient leaves against medical advice, and later is readmitted, that readmission should not count for analysis since they do not affect the discharge decision. The overall goal of my research is to improve physician discharge decision for cardiac patients, and such discharge would be inappropriate to include for analysis of readmission patterns. Also, other Discharge Disposition is considered non-acute medical services and are not associated with the discharge decision like “Rehab facility” and “Psychiatric facility.” These are visits that were listed as readmission but do not affect improving the discharge decision. I used an R script to clean the queried data from issues found in the preliminary work like null values, unordered visits per patient, and duplicate values. Also, one of the steps in getting clean data is to look for discharges that were recorded due to the billing procedure to mark the end of the service before starting the new service. These two services could have been performed in the same unit or a different unit. For these reasons, I made sure that the code understands that pattern and automatically mark them as a transfer. Our methods also included removing data that were marked as discharge but should be more appropriately viewed as a transfer. For example, sometimes a patient may be transferred between units within the hospital, yet these transfers are documented as a discharge, followed by a re-admission. Again, this sort of readmission should not be included in a visit if the goal is improved decision support.

3.2 Data Quality Assessment

After creating the visit table, I wanted to assess the quality of the visit table through the following stages:


Compare it to UW Medicine Annual Report
Compare it to the Literature
Patient Discharge Summary report
Consult Cardiologist
Patient Chart Review

I started by comparing my readmission number with UW Medicine Annual Report. The UW Medicine Annual Report is a detailed statistical report that UW Medicine generates every year about the number of inpatient and outpatients for a different diagnosis. For that, I selected a specific cohort that matches the cohort in the report (For example, 2016 only). (Figure 3.5) However, when I controlled for only 2016 visits, I get 44,330 admission/visits from University of Washington Medical Center (UWMC), Harborview Medical Center (HMC), and Northwest Hospital and Medical Center (NWH). However, the report shows that admission from these three centers in 2016 is 45,391. There is a discrepancy of about 1000 visits/admission. Different reasons could cause such discrepancy and with the high uncertainty of how the report has been pulled and what control has been used, makes it hard to investigate further.



FIVE-YEAR PERFORMANCE COMPARISON

Northwest Hospital & Medical Center

Statistic	FY 2012	FY 2013	FY 2014	FY 2015	FY 2016
Admissions	9,127	9,974	9,211	9,934	10,060
Patient Days	43,350	44,333	44,189	47,143	48,492
Outpatient Visits	193,992	195,978	193,387	195,031	197,132
Emergency Visits	33,832	33,942	34,276	36,159	35,068
Average Length of Stay	4.7 days	4.4 days	4.8 days	4.7 days	4.8 days
Case Mix Index (CMI)	1.37	1.42	1.43	1.46	1.50

University of Washington Medical Center

Statistic	FY 2012	FY 2013	FY 2014	FY 2015	FY 2016
Admissions	17,915	17,728	18,033	18,092	18,362
Patient Days	120,745	122,867	124,513	126,239	132,529
Outpatient Visits	300,487	284,870	291,375	302,038	320,037
Emergency Visits	23,487	22,977	25,338	26,465	26,555
Average Length of Stay	6.7 days	6.9 days	6.9 days	7.0 days	7.2 days
Case Mix Index (CMI)	1.99	1.98	2.02	2.13	2.24

Harborview Medical Center

Statistic	FY 2012	FY 2013	FY 2014	FY 2015	FY 2016
Admissions	19,094	17,999	17,176	17,362	16,969
Patient Days	134,930	135,779	132,284	138,214	144,140
Outpatient Visits	244,964	245,751	247,349	247,615	252,435
Emergency Visits	62,432	66,285	64,512	62,217	59,776
Average Length of Stay	7.1 days	7.5 days	7.7 days	8.0 days	8.5 days
Case Mix Index (CMI)	1.91	1.99	2.10	2.15	2.23

Figure 3.5: UW Medicine Board Annual Financial Report Five Year Performance

Then, I compared the number I got from my analysis with results from some literature. I started with a study done by Eastwood et al. 2014 where they compared seven days readmission with 30 days readmission. Eastwood study has looked up the determinant factors that would increase the likelihood of early readmission, focusing mainly on 7 and 30 days readmission. In that study, they found that 5.6% of their 18,560 HF patients were readmitted within seven days and 18% were readmitted within 30 days. The study has a broader definition of heart failure as it compasses the main ICD 10 class of HF I50.x and all its subclasses. The study has neither looked into the cardiac severity such as ejection fraction nor the type of medication used inside the hospital. In my study, we have a total of 2032 HF patients, and 7-Days readmissions happen for 16.1% of our HF cohort. The different of HF definition among studies in the literature makes it difficult to be compared to for quality assessment.

Also, a cardiologist and I have gone through some of the patient discharge summary report and patient discharge summary report for small population looking into readmissions within 24 hours. Patient chart review can be a time-consuming process. I wanted to know how readmission, transfer, and discharge is defined in the EHR on a small cohort. When creating the visit table, I need to differentiate readmission from transfers despite the billing procedure. I need to distinguish between transfers and readmissions and know what count as a readmission. For example, in the previous analysis, we found that for some heart failure patients they marked their visits to physical therapy and physiology as readmission instead of transfer. Such visits have been excluded from the analysis as they are not considered as an acute admission. It is important to have an accurate visit table as the retrospective data from the visit table will be the input for creating a readmission prediction model.

3.3 RESULTS

After querying the visit data from Caradigm, there are 260776 visits from 158695 patients from 2009 to the beginning of 2018 (queried on 1/25/2018). Then, after applying our heart failure definition, I ended up with 1385 patients that have a diagnosis of heart failure and all-cause readmission after that. These 1385 patients resulted in 7194 separate visits. Then, I delete the initial diagnosis since I am looking to analyze readmission instances only. This resulted in 5488 visits from 1123 patients. From those, 2030 visits from 688 heart failure patients happened within 30 days (~%30). The process described below is my detailed data cleaning process of coming up with the numbers for the visit table

I started manually analyzing readmission within the first 24 hours of discharge. This will make the manual check easier since it will be on a small population. I noticed that some of these visits have Account Status that says “Discharged” and DemoDischargeDispositionDescr: “HOME: Home/Self Care,” yet the admit source description says, “Transfer w/in Hospital Resulting in Separate Claim to Payer.” These are mostly transfers either from the operating room or the emergency department but for billing reasons they mark them as discharged to home. So, I merged the multiple admission for that patient into one. So, I eliminated rows that its DemoDischargeDispositionDescr says "TRANSFER TO: Transfer to Hospital" and its AdmitSourceDescr says "Transfer w/in Hospital Resulting in Separate Claim to Payer." I eliminated patients who left against medical advice since it does not affect the discharge decision. Then, I noticed also that some of the readmission are for services that are not considered acute medical services such as Psychiatry and Physical Medicine and Rehabilitation should not have readmission counted if admission is to one of these services. For that, I eliminated readmission that is considered Psych or Rehab. For most of the steps above, I take the patient ID number and manually check their readmission pattern from the main visit table. Finally, I applied the same restrictions from the 24-hour readmission to the 30-days readmission. This resulted in 688 patients and 2030 visits. Figure 3.6 below shows the distribution of heart failure patient readmission within 30 days. The histogram shows that ~40% of the 30-day readmissions happen within the first seven days after the patient has previously been discharged.

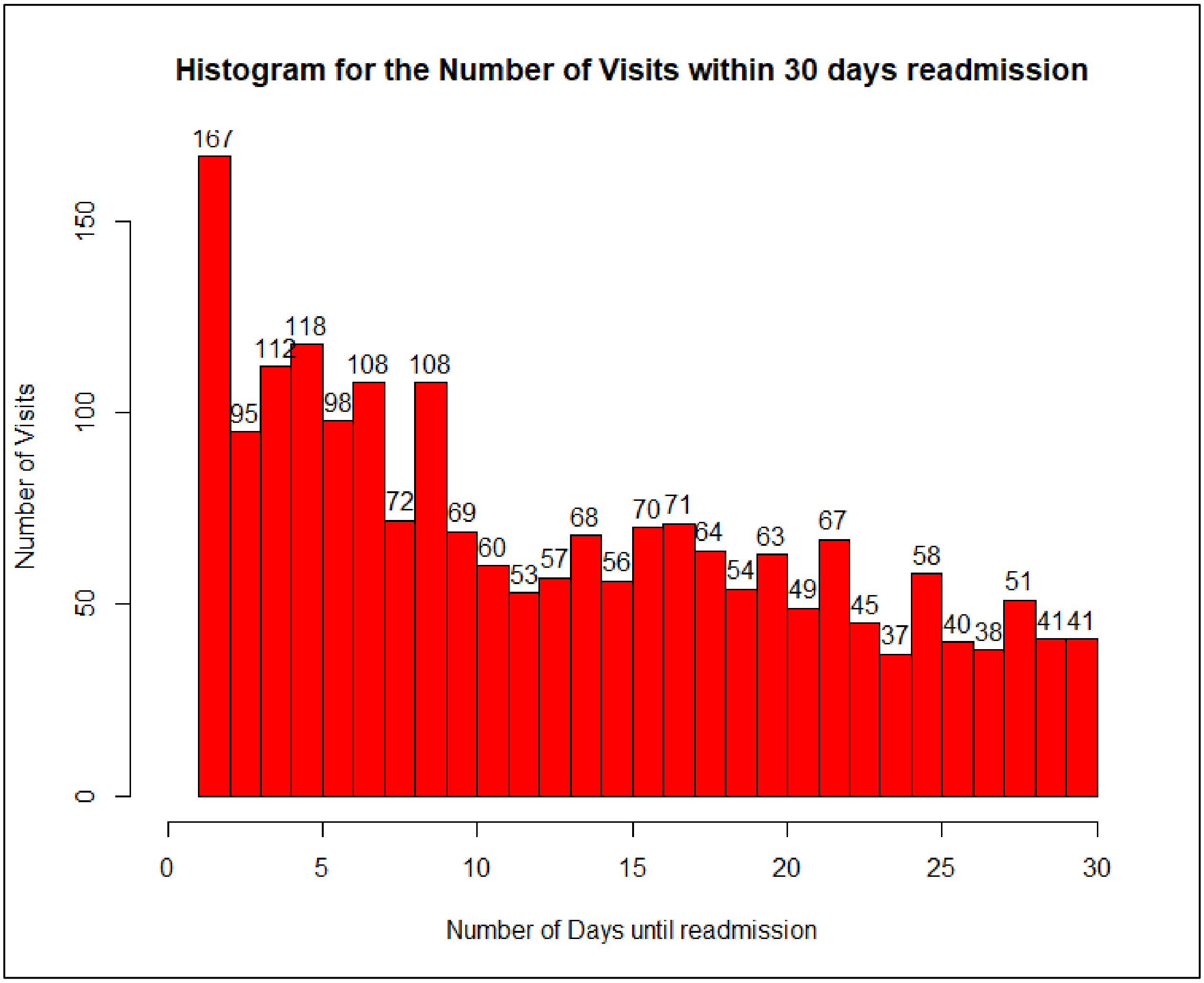


Figure 3.6: Histogram for the Distribution of Readmission within 30 days

Reason for Deletion	Visits Deleted
Transfer to Hospital	31
Left Against Medical Advice	264
Disch/Trans to a Distinct Psychology Unit	30
Transfer to Different Institute (not specified)	65
Transfer within Hospital	14
Transfer to REHAB	9

Table 3.3: The changes in the number of visits after deleting certain services when counting readmission. The original number of visits for Heart Failure Patients is 7194 visits.

As table 3.3 shows, my methods resulted in approximately a 6% reduction in data. After cleaning, we assessed the accuracy of our visit table via consultation with a cardiologist and manual chart reviews. Although not a very large percentage of data, this cleaning step could be critical to analysis, as it may be that the data cleaned are outliers, and might, therefore, have a big impact on aggregate summaries that would be important for analyses.

3.4 DISCUSSION

My goal is to build a visit table that I can use to build prediction models for heart failure patients at risk of readmission within 30 days. More generally, I seek to understand which of my steps and methods for creating the visit table can be generalized to other institutions. To the extent that they generalize, I could build a tool that would help researchers automate the process of building an appropriate visit table, and reduce the need for time-consuming chart review. This tool would produce a “cleaned” visit table, which could then be the nucleus for some analyses, including prediction tasks.

Many of the data noise could be figured out by researchers such as null values, unordered visits, redundancy, and transfers versus readmission. However, some of the data issues require an expert in the field to help determine what count as readmission and what is not. This can be seen clearly in differentiating between acute readmission and non-acute readmission. Also, what discharge disposition need to be included in analyzing patient readmission. The need to automate this process is important to empower researchers to be more independent as consulting cardiologist is not accessible for everyone.

Chapter 4. CHARACTERIZING DAY 1-30 READMISSIONS

In this chapter, I will characterize patient readmission patterns using the visit table created in chapter 3. There have been many efforts made by researchers to predict the risk of readmission within 30, 60, 90, 180 days and one year of discharge. More specifically, researchers have exhaustively investigated reducing readmission within 30 days as a measure of the quality of care in hospitals. Now, hospitals will be penalized with worse than 30 days readmission as part of the affordable care act. In chapter 2, I looked into the previous efforts in predicting heart failure readmission and finding the patient characteristics that are associated with a higher risk of readmission and/or death. In my research, I focused on models that either predicts heart failure readmission alone or readmission and death, excluding models that use death as their only output measure.

As I discussed in chapter 2, my work on selecting models to compare to is based on two systematic reviews about readmission prediction models for heart failure patients; Ross et al. (2008) and a recent study by O’Connor et al. (2016). (Ross, Mulvey, Stauffer, & al., 2008) (O'Connor, et al., 2016) Also, it is based on two registries; the Acute Decompensated Heart Failure National Registry (Adhere) and the Organized Program to Initiate Lifesaving Treatment in Hospitalized Patients with Heart Failure (OPTIMIZE-HF), which were built to encourage the use of evidence-based research to bridge the gap between knowledge and medical care.

In my study, I want to investigate when patient characteristics changes during 1-30 days all-cause readmission for the heart failure patient. For example, is there a significant difference in the characteristics of 4-day readmission patients versus 24-day readmission patients? To answer this question, using an accurate visit table, I will build a prediction model based on retrospective data created from the visit table. The idea is to conduct the analyses on a retrospective data on five days ranges for the 30 days readmission patients to find where the cutoff in patient characteristics exists. In other words, on what day of readmission within the 30-day window do we see changes in patient characteristics?

The goal of my work is to predict “very early re-admission” and try to prevent them by improving the decision of discharge. I hypothesize that “very early re-admission” are those readmissions that are more likely due to error, or readmissions, that if avoided, could have led to substantial cost savings and/or potentially improved health outcomes.

As I described in chapter 2, both Ross et al. (2008) and O’Connor et al. (2016) systematic review yielded no consistent predictors of readmission among the different models they analyzed. (Ross, Mulvey, Stauffer, & al., 2008) Also, they mostly looked into 30 days readmission and higher ranges. As I described in chapter 2, Eastwood et al. (2014) was the only study I could find that looked at shorter than 30 days readmission for heart failure patients where the study looked into 7-day readmission. They have looked into determining the factors that could increase the risk of readmission within seven days and 30 days. (Eastwood et al., 2014). The study concludes that these seven days readmission is associated with the history of kidney disease. The study has neither looked into the cardiac severity such as ejection fraction nor the type of medication used inside the hospital or Echo data. I believe that measuring the severity of the disease and factoring the medication intake inside the hospital is vital for predicting readmission, which I will include in my model. Moreover, evidence-based research that utilized the Adhere and OPTIMIZE registries have mostly looked into 30 days readmission.

I need to define the predictors associated with readmission for the heart failure patient. Some of these variables exist in the current extracted dataset, and some are not that need to be queried separately. These predictors include but not limited to; Demographics, Medical History, Laboratory Values, Weight…etc.

The visit table created in chapter 3 include some demographic information such as age and gender; it also has information about medical history and heart failure history. However, some of these variables are documented in free text format which will require me to apply some natural language processing techniques to extract the unstructured data. Another resource about the medical history of admission can be found in the Diagnosis table. However, there are some variables that I was not able to locate in the current dataset such as ECHO data. For variables with different measures during the visit such as Heart Rate, Ejection Fraction, and Weight, I will use the last number before they got discharged.

To build the prediction model, I will start by replicating both the Adhere and Optimize models using my visit table I create from Caradigm data. The test will be for predicting 30 days readmission. I will start by using the same predictor's variables they used in building their prediction model. Then, I will add variables that I believe are important but were not included in the Adhere and OPTIMIZE model such as vital at discharge, in hospital medication, length of stay in the hospital, and the number of previous readmissions. The goal is to test if adding these variables will improve the accuracy of the model. Then, the last model to run is with adding Echocardiography data with all variables and see if ECHO data can improve the accuracy of the model. Finally, I will choose the model that will provide the highest prediction accuracy on predicting 30-days readmission to conduct the 6-days range characterization.

Windows of days	Number of Visits
1-6	590
2-7	634
3-8	603
4-9	616
5-10	573
6-11	515
7-12	470
8-13	419
9-14	415
10-15	363
11-16	364
12-17	375
13-18	386
14-19	383
15-20	378
16-21	371
17-22	368
18-23	342
19-24	315
20-25	319
21-26	296
22-27	285
23-28	269
24-29	265
25-30	269

Table 4.1: Windows of days and their number of visits

I will base my analysis on visits characteristics, not patients, as patient characteristics might change from visit to another and I am analyzing visits characteristics at the time of discharge to predict very early readmission. In defining very early readmission, I will conduct my analysis on a slider range of 6 days (that is 1-6, 2-7, 3-8…etc.) using Logistic Regression. For example, I will compare patients who are readmitted on days 2-7 and compare them to patients who got readmitted on days 3-8. However, I switched to a larger window (that is 7-days) when the number of visits started to drop after the 9th day. (see table 4.1) I will compare positive examples (people who got readmitted within the specific range) with people who never got readmitted within one year. In splitting the data into train and test, I will be using Cross-validation.

4.1 DEFINING THE VARIABLES (DATA PREPARATION)

The visit table I created in the previous chapter and the data cleaning that I conducted in the process does not make the variables ready to plug in into the model. There are some variables that I will need to create from the visit table and some I need to organize, and others need to be queried and added to the visit table. An example of variables that needed to be created from the visit table is Medical History. Patient medical history is variables that are buried under seven columns. These seven columns take the form of either free text data or ICD codes that describe the disease. These seven columns are the following; the admit reason, primary diagnosis, primary diagnosis ICD code, secondary diagnosis, secondary diagnosis ICD code, tertiary diagnosis, and tertiary diagnosis ICD code. In my R code, I used the function ‘sapply’ to search the free text vector for different naming of the diseases under the columns using ‘grepl’ function to return TRUE if a string contains the pattern. For example, Stroke/ Transient Ischemic Attack is written differently for each visit (row). It is documented in the following different formats; Stroke,

stroke, STROKE, transient ischemic attack, or TRANSIENT ISCHEMIC ATTACK. The code needs to be able to locate the different ways the disease was input in the EHR. Moreover, the ICD codes of the disease is also another way to mark the history of the disease since for each diagnosis; primary, secondary, and tertiary, there is an ICD code column. The University of Washington Caradigm contain visits that coded with ICD-9 and ICD-10 code. It is important to include the ICD-9 and ICD-10 codes for each disease when searching the dataset. (See table 4.2)

Medical History	**ICD 9**	**ICD 10**
Atrial Fibrillation	427.31	I48.0
Coronary Artery Disease	414, 414.01, 414.02, 414.06	I25.10, I25.110, I25.118, I25.119, I25.810, II25.811, I25.82
Congestion	514	-
Chronic Obstructive Pulmonary Disease	491.21	J44.0, J44.1, J44.9
Chronic Renal Insufficiency	403.9, 403.91, 404.91, 404.93, 585.2, 585.3, 585.4, 585.6, 585.9, 586	I12.0, I12.9, I13.0, I13.2, N18.4, N18.9, N18.3
Diabetes	250, 250.4, 250.6, 250.8	E11.21, E11.22, E11.319, E11.359, E11.40, E11.32, E11.43, E11.51, E11.52, E11.611, E11.621, E11.628, E11.641, E11.649, E11.65, E11.69, E11.9
Fatigue	780.7, 780.79	R53.8
Hyperlipidemia	272, 272.4	E75.5
Hypertension	401, 401.9, 405.91, 416, 572.3	I10, I1.8, I27.0, I27.2, I27.20, K76.6, 010.911, I97.3
Peripheral edema	782.3	R60.9

Peripheral vascular disease	443.9	I73.9
Rales	786.7	R09.89
Stroke/transient ischemic attack	433.10, 434.11, 434.9, 435.9	I63.9, G45.9 Z86.73
Ventricular tachycardia/ventricular fibrillation	427, 427.1, 427.41	I47.1, I47.2, I49.01
Myocardial Infraction	410.9	I21.3, I21.4, I21.9, I22

Table 4.2: Medical History ICD 9 and ICD 10

Another variable that needs to be prepared for the model is the Left Ventricular Ejection Fraction (LVEF). Ejection Fraction is a measurement, expressed as a percentage, that represents how much the left ventricle pumps out blood with each contraction. A 30 percent Ejection Fraction means that only 30 percent of the total amount of blood is pushed out with each heartbeat. Based on the American Heart Association, a normal Ejection Fraction is said to be between 50-70 percent. An ejection fraction under 40 percent may be evidence of heart failure. An ejection fraction of higher than 75 percent could indicate a heart condition such as hypertrophic cardiomyopathy. ("Ejection Fraction Heart Failure Measurement | American Heart Association," n.d.) LVEF is an important indicator of heart failure patients risk of readmission. (Loop et al. 2016). Ejection Fraction data exist in a different dataset. That dataset contains Patient ID, Visit ID, Study Date, LEVF quantitative variable, Finding comments, machine name. I merge that dataset with my visit table by the visit ID and add the LEVF quantitative variable as a column in my heart failure visit table.

In Hospital Medication:

I hypothesize that in-hospital medication for patients with prior diagnosis of heart failure can help predict the risk of readmission. In-hospital medication has been queried in a different dataset with detailed information about medication name, medication route, dosage, and date and time of medication. After merging this dataset with my visit table, there are 31 different medications listed under the different visits. Adding all 31 as variables would increase the dimensionality of the model. For that, I classified these medications based on their medication class using RxNorm and then approved the list with a cardiologist. RxNorm is a tool produced by the National Library of Medicine (NLM) that support semantic interoperation between the pharmacy knowledge base and drug terminologies. It provides normalized names for drugs and links them to many of the drug vocabularies. ("RxNorm," n.d.) Figure 4.1 shows an example from RxNorm of drug classification. The classification of the 31 drugs yielded into eight different classes. These classes are; Statins, Diuretics, Beta Blockers, Angiotensin Converting Enzyme (ACE), Angiotensin Receptor Blocker, Aldosterone Antagonist, Nitrates, and Cardiac Glycosides. After consulting a cardiologist, I was asked to remove three medications due to their insignificance to the study purpose. (see table 4.3)

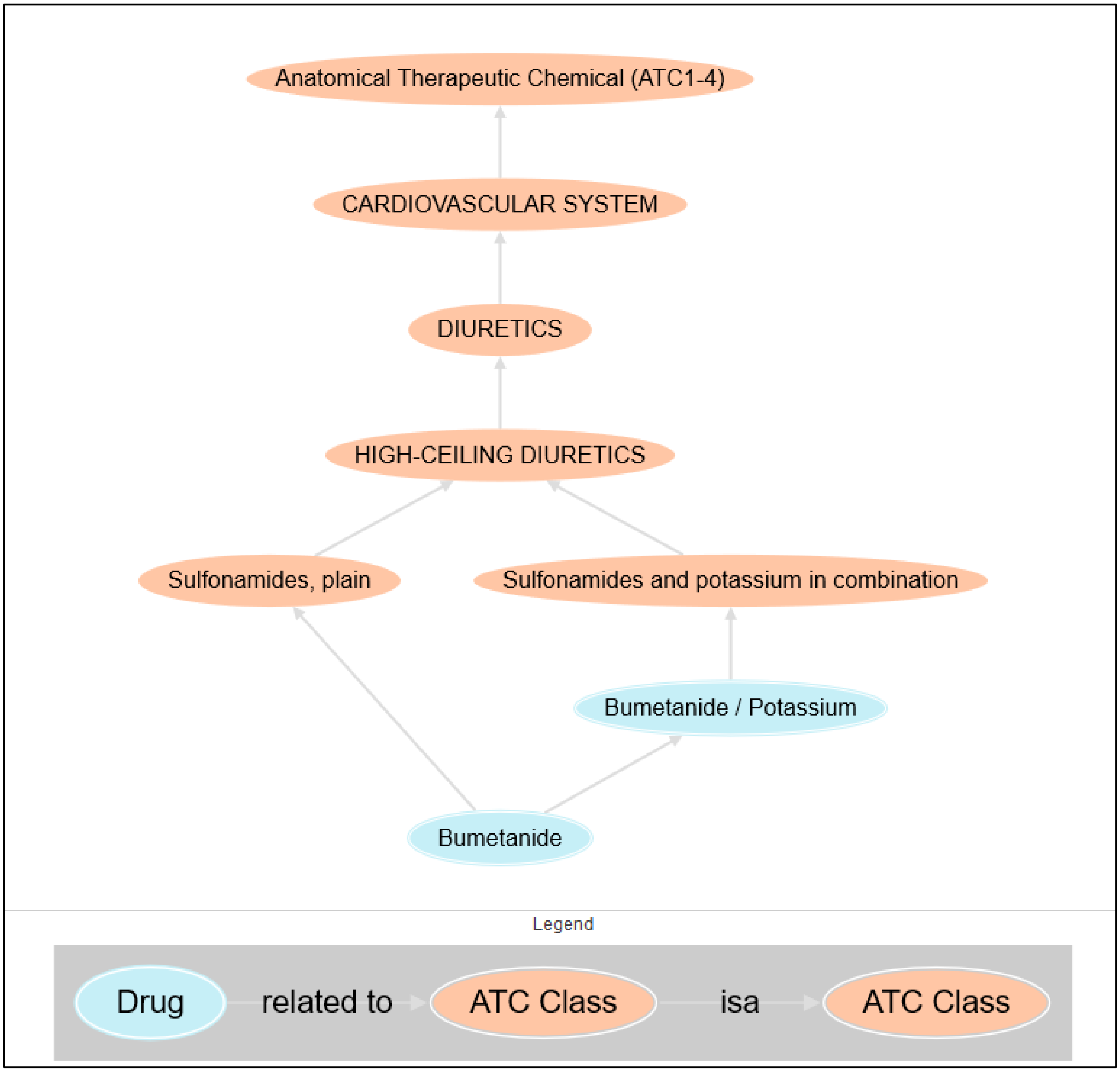


Figure 4.1: An RxNorm Example of Drug Classification (RxNorm website)

No.	Medication Name	Medication Category
1	Atorvastatin	Statins
2	bisoprolol	Beta Blockers
3	bumetanide	Diuretics
4	candesartan	Angiotensin receptor blocker
5	CAPtopril	ACE

6	chlorthalidone	Diuretics
7	digoxin	cardiac glycosides
8	enalapril	ACE
9	eplerenone	Aldosterone Antagonist
10	ezetimibe-simvastatin	Statins
11	Fluvastatin	Statins
12	furosemide	Diuretics
13	hydrochlorothiazide-spironolactone	Aldosterone Antagonist
14	irbesartan	Angiotensin receptor blocker
15	isosorbide dinitrate	Nitrates
16	isosorbide mononitrate	Nitrates
17	lisinopril	ACE
18	losartan	Angiotensin receptor blocker
19	lovastatin	Statins
20	metolazone	Diuretics
21	metoprolol	Beta Blockers
22	nebivolol	Beta Blockers
23	pitavastatin	Statins
24	pravastatin	Statins
25	ramipril	ACE
26	rosuvastatin	Statins
27	sacubitril-valsartan	Angiotensin receptor blocker
28	simvastatin	Statins
29	spironolactone	Aldosterone Antagonist
30	torsemide	Diuretics
31	valsartan	Angiotensin receptor blocker

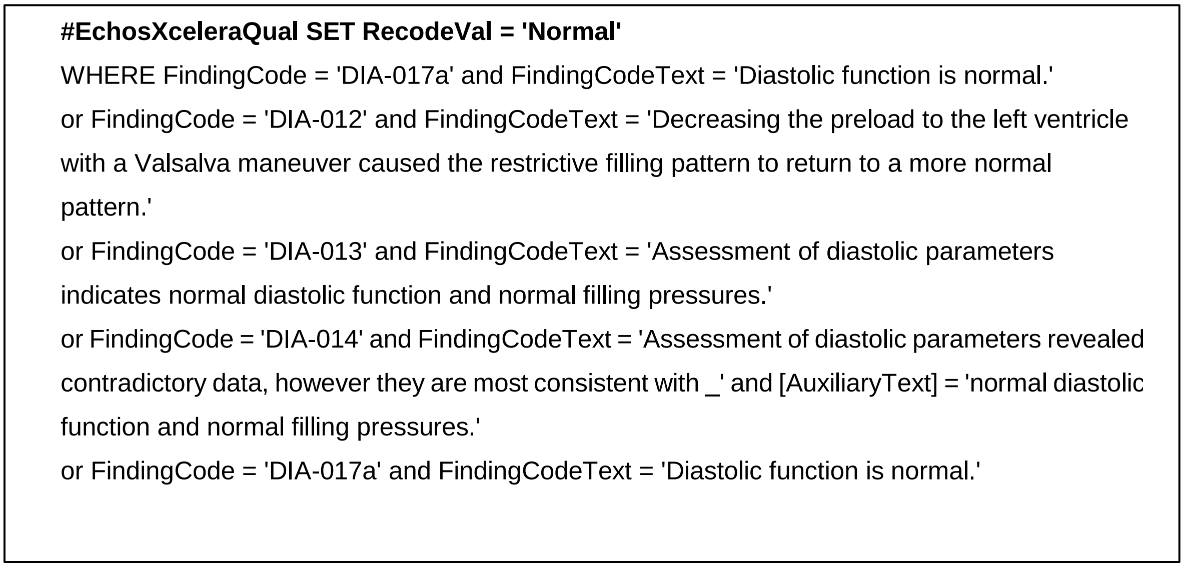
Table 4.3: In Hospital Medications and their Classification

Race:

The race is another variable that needs to be prepared and properly cleaned. Unfortunately, race in the EHR is not standardized which yielded different terms when coding a race. This will increase the dimensionality of the model and decrease the relevance of the race variable to the model. For example, Chinese, Flipino, Cambodian, Vietnamese, Japanese, Korean, and Thai are all different examples of the Asian race. Moreover, Black, African American, Black/African American are different ways that the EHR uses to represent one race. For my model, I deduce the 23 different formats of race in the EHR into nine different types of race or variables using The United States Census Bureau race grouping. ("US Census Bureau," n.d.) The nine races variables are; Caucasian, Black or African American, Asian, Native Hawaiian or Other Pacific Islander, American Indian or Alaska Native, Hispanic, Multi-Racial, Other, Not Reported or Unknown, and Declined to answer.

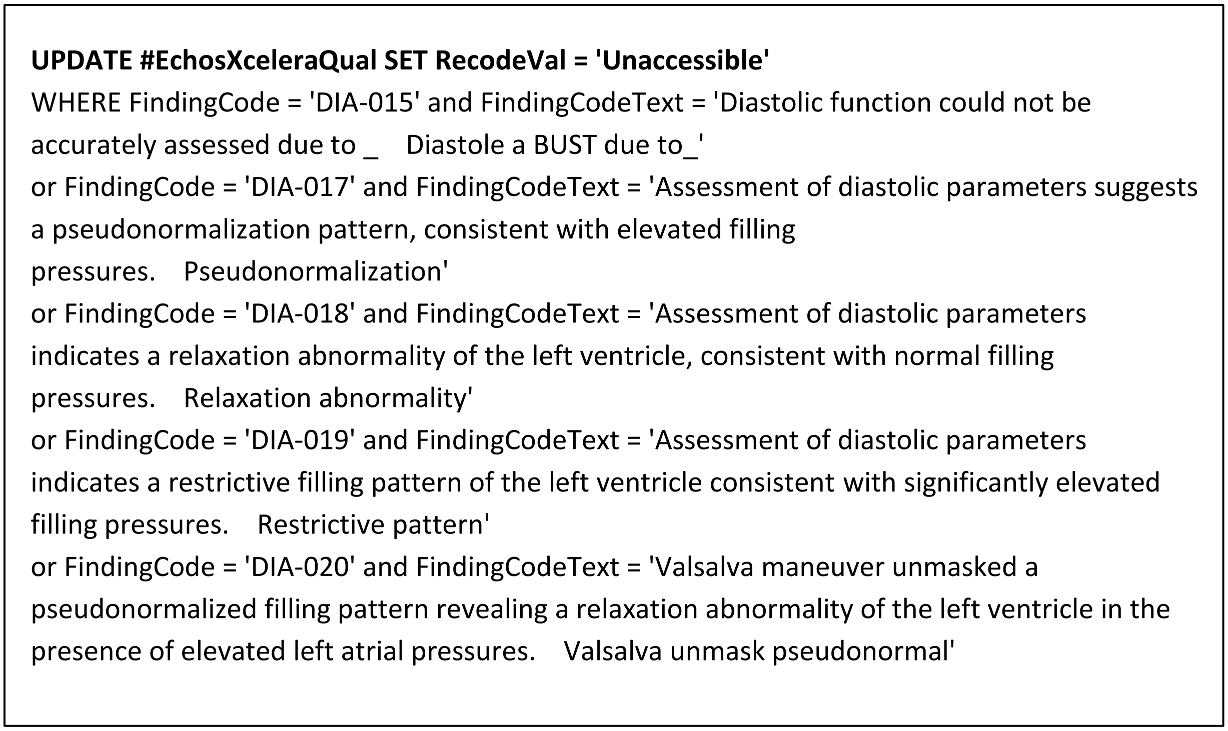
Echocardiography data:

Another important variable that I believe will improve the prediction of my model and has not been used in previous studies is Echocardiography data. Echocardiography or echo test is a test that takes "moving pictures" of the heart with sound waves. (American Heart Association., 2013) The test provides data about heart structure and performance. Analyzing the Echo test could be very complicated as it contains dense qualitative and quantitative information about the structure of the heart. Each echo test in the dataset, queried from Caradigm, will have a quantitative report and qualitative findings. These reports and findings were recorded into Normal, Abnormal, and Unaccessible values using logic given by a cardiologist. Figure 4.2 and 4.3 are examples of how the echo test was coded. For my study purpose, I will use the record value generated from the cardiologist logic. The variables that will be added to my model are Normal, Abnormal, Unaccessible Echo data.



```
#EchosXceleraQual SET RecodeVal = 'Normal'
WHERE FindingCode = 'DIA-017a' and FindingCodeText = 'Diastolic function is normal.'
or FindingCode = 'DIA-012' and FindingCodeText = 'Decreasing the preload to the left ventricle
with a Valsalva maneuver caused the restrictive filling pattern to return to a more normal
pattern.'
or FindingCode = 'DIA-013' and FindingCodeText = 'Assessment of diastolic parameters
indicates normal diastolic function and normal filling pressures.'
or FindingCode = 'DIA-014' and FindingCodeText = 'Assessment of diastolic parameters revealed
contradictory data, however they are most consistent with _' and [AuxiliaryText] = 'normal diastolic
function and normal filling pressures.'
or FindingCode = 'DIA-017a' and FindingCodeText = 'Diastolic function is normal.'
```

Figure 4.2: An Example of the Logic Used to Code Echo Results (Normal)



```
UPDATE #EchosXceleraQual SET RecodeVal = 'Unaccessible'
WHERE FindingCode = 'DIA-015' and FindingCodeText = 'Diastolic function could not be
accurately assessed due to _    Diastole a BUST due to_'
or FindingCode = 'DIA-017' and FindingCodeText = 'Assessment of diastolic parameters suggests
a pseudonormalization pattern, consistent with elevated filling
pressures.    Pseudonormalization'
or FindingCode = 'DIA-018' and FindingCodeText = 'Assessment of diastolic parameters
indicates a relaxation abnormality of the left ventricle, consistent with normal filling
pressures.    Relaxation abnormality'
or FindingCode = 'DIA-019' and FindingCodeText = 'Assessment of diastolic parameters
indicates a restrictive filling pattern of the left ventricle consistent with significantly elevated
filling pressures.    Restrictive pattern'
or FindingCode = 'DIA-020' and FindingCodeText = 'Valsalva maneuver unmasked a
pseudonormalized filling pattern revealing a relaxation abnormality of the left ventricle in the
presence of elevated left atrial pressures.    Valsalva unmask pseudonormal'
```

Figure 4.3: An Example of the Logic Used to Code Echo Results (Unaccessible)

Age:

During my analysis, I have noticed that the Age variable was missing in about ~17% of the total visits. (See Figure 4.4) It is clear that something happened when the data was queried. Since I am working with a dataset that was queried from Caradigm and my IRB only permits me to work with the queried dataset, I have no access to Caradigm data to check for such an error. Moreover, the queried dataset does not include the patient date of birth to help me calculate their age. I assume that the date of birth was removed for privacy reason since patient and Visit ID were hashed as well. However, I can fix most of the missing data by calculating the birth year of the patient by subtracting their age from a previous visit from the admission date of that year. I cross checked this method with visits where age was not missing, and the result was checked out to be accurate. This method improved the missing age data to be just (~3%). These %3 missing data is from patients where age was missing in all their visits.

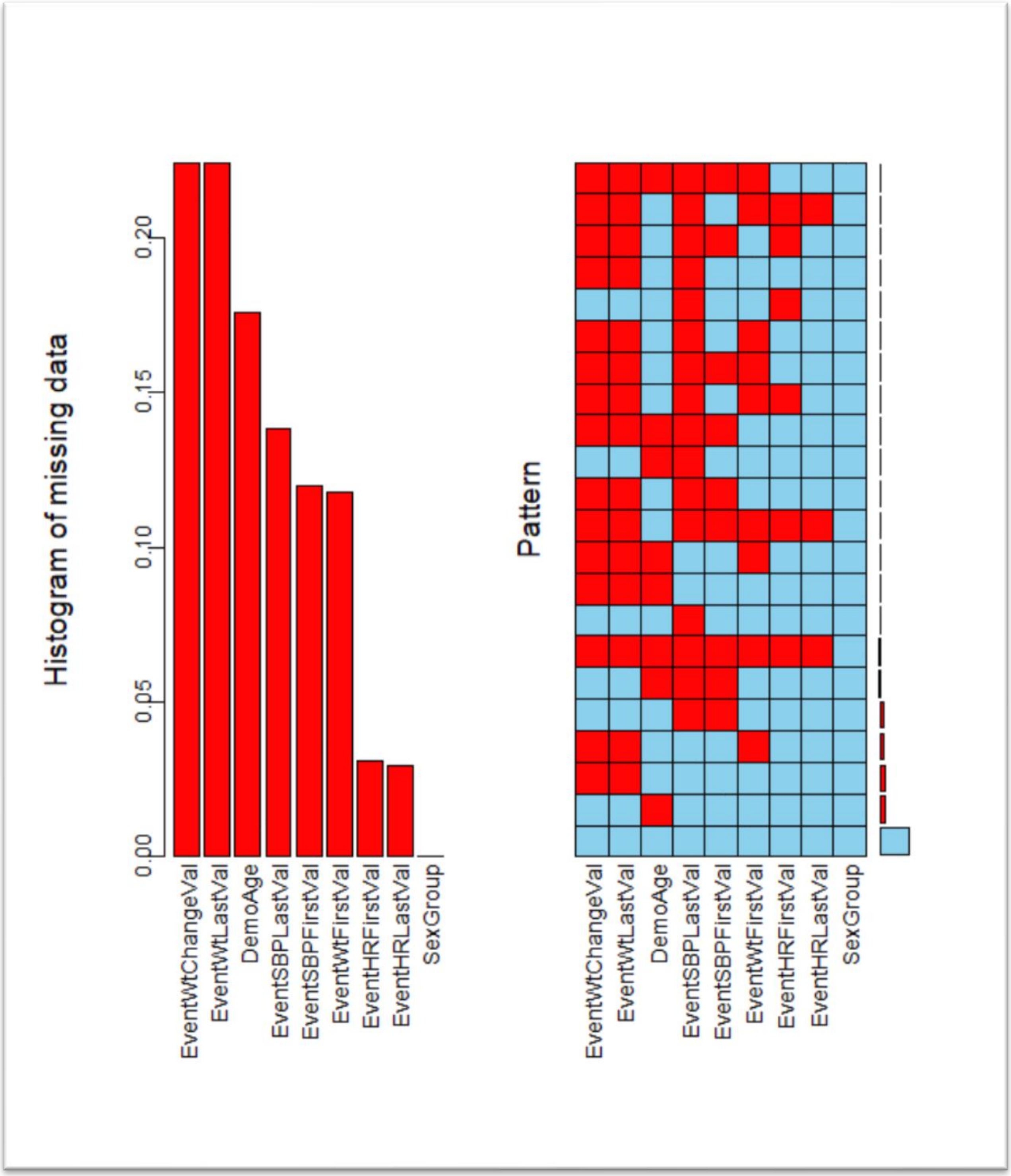


Figure 4.4: Histogram of Missing Variables

4.2 MODEL SET UP

After the data cleaning and preparation, there are a total of 58 variables in the heart failure patients visit table. These variables vary from demographics, primary insurance, medical history, laboratory variables, and vital signs. The original visit table contains 7194 visits from 1385 heart failure patients. However, since my study focuses mainly on readmission, I dropped the initial patient visit. Also, I dropped patients with only one initial visit and no readmission. This yielded to 5488 visits from 1123 patients. From those, 2030 visits from 688 heart failure patients happened within 30 days (~%30). The 30 readmission visits pattern can be seen in Figure 4.5.

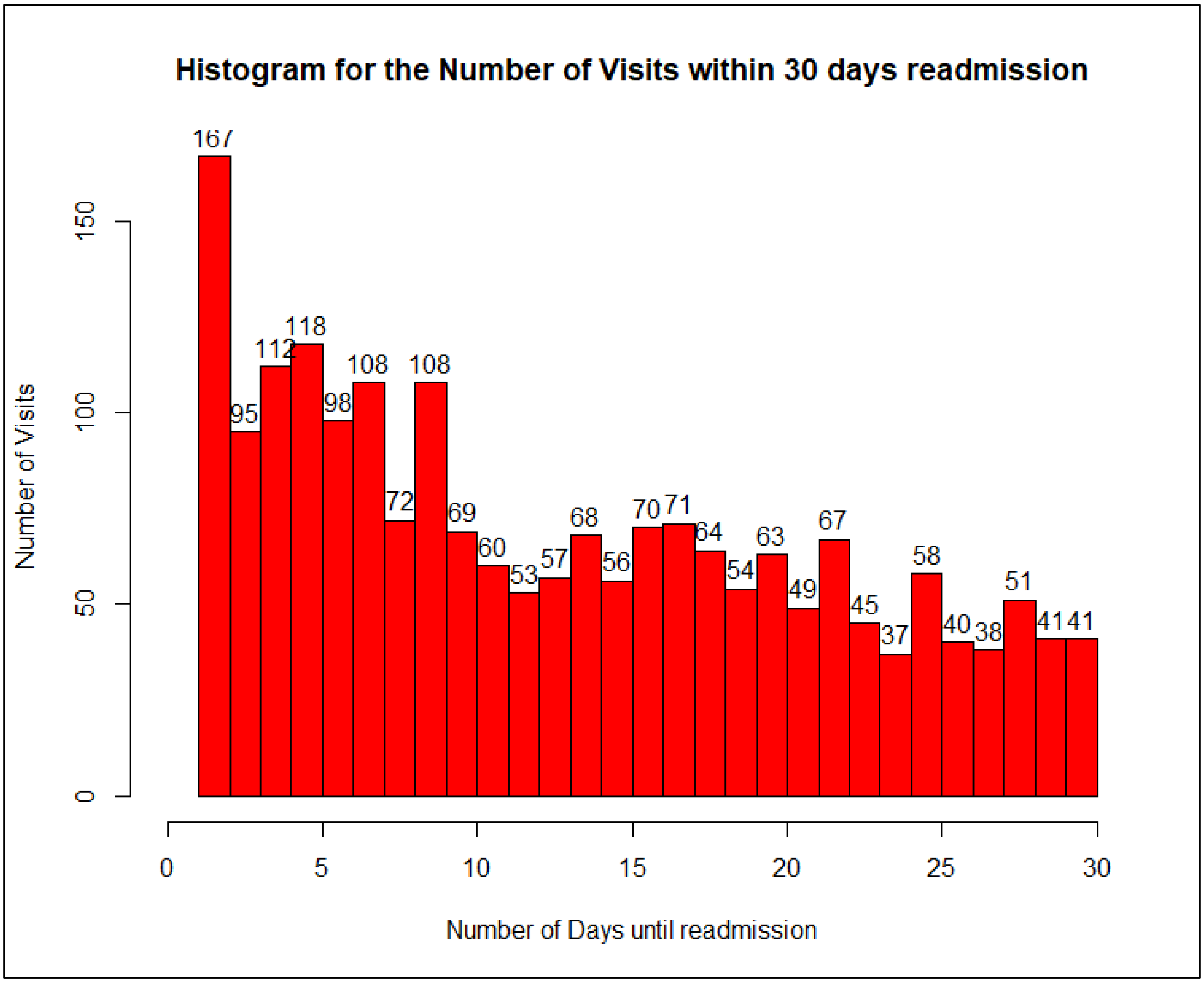


Figure 4.5: Histogram for the Number of Visits Within 30 Days of Readmission

Outcome Variable:

The primary objective of my model is to improve discharge decision making by predicting early readmissions before discharge. Therefore, the outcome variable of my study is a binary outcome where I am checking if patients got readmitted within a specific range of days or not. As mentioned in chapter 3, I am analyzing prospective data of all-cause readmission for heart failure patients, from UW Medicine EHR Data, for readmissions from 2010 till the end of 2017. In this analysis, I started by analyzing 30 readmissions on three models; Adhere model variables, Adhere model variables with medication and vitals at discharge, and finally adding Echo data. The goal is to locate the model with higher prediction score. Then, after choosing the model with the highest prediction score, I conduct my analysis based on six days until readmission increment. This means that I build a model for day 1-6, 2-7, 3-8, 4-9…etc of readmission. Once I found the day where the visit characteristics started to change, then I will define that range of days as the very early readmission range.

Statistical Methods:

I continued building on the R code that I used for creating the visit table. In this analysis, I have used a Logistic Regression model for predicting if the patient will be readmitted within the predefined readmission range. I started with the full logistic regression model including all variables to identify significant factors contributing to readmission. Then, I used the step() function in R to automatically delete all the insignificant variables setting the direction to be "both." This means that we start with the full model, then we consecutively "both" remove insignificant variables and also recruit new ones. Then, I conduct an ANOVA test between the two models the full and reduced, to test if the two models are statistically different. The ANOVA function reports a p-value via the Likelihood Ratio Test (LRT). The LRT test expresses how many times more likely that data are under one model than the other. (Murphy, 2012). If there are no statistical differences between the two models, I will reduce the reduced model since it performs as good as the full model. I have split the data into train and test data using 10 fold Cross-Validation. My strategy in creating the different five days ranges models is by finding the visits that happened within the defined range and then I randomly select visits that happened after 30 days but within one year. The reason I made the cutoff to be within one year is that we currently do not have a record in the EHR to tell us if the patient died at home. We only have a record of patient death if it happens in the hospital. For that reason, leaving the range of readmission open makes it hard to distinguish if the no readmission within specific date happened because the patient is healthy or merely that they died.

Variables	Adhere	My Model
Demographics		
Age	X	X
Sex	X	X
Height	X	
Weight	X	X
Race	X	X
Primary Insurance	X	X
Heart Failure History		
Prehospital	X	
Ischemic etiology	X	X
Baseline NYHA class	X	(4 cells only, free text)
NYHA class at presentation	X	
Medical History		
Coronary artery disease	X	X
Prior myocardial infarction	X	X
Prior revascularization	X	

Atrial fibrillation	X	X
Congestion	X	X
Chronic obstructive pulmonary disease	X	X
Chronic renal insufficiency	X	X
Diabetes	X	X
Duration of symptoms	X	
Fatigue	X	X
Hyperlipidemia	X	X
Hypertension	X	X
Peripheral edema	X	X
Rales	X	X
Stroke/transient ischemic attack	X	X
Ventricular tachycardia/ventricular fibrillation	X	X
Laboratory values		
B-type natriuretic peptide	X	X
Blood urea nitrogen	X	X
Cardiac enzymes	X	
Creatinine	X	X
Dyspnea at rest	X	
Hemoglobin	X	X
QRS duration >120 ms	X	
Qualitative LVEF	X	X
Sodium	X	X
Initial vital signs		
Diastolic blood pressure	X	
Systolic blood pressure	X	X
Heart rate	X	X

Table 4.4: Variables Comparison between Adhere and my Model

4.3 RESULTS

The baseline characteristics of the 5488 visits from 1123 patients of the study cohort are shown in table 4.5. I have started my analysis comparing my model to the Adhere Model by using the same variables. (see table 4.4) Of the 46 variables evaluated for the risk of 30 readmissions, Systolic Blood Pressure (SBP) at admission is the most predictable variable with a p-value of 0.013. The second most significant variable is Medicaid Insurance with a p-value of 0.06. Mimicking the Adhere Model variables resulted in 0.64 C-Statistic score.

Then, I added to the 39 variables the following variables; vital at discharge, In-hospital medication, Length of Stay, and Number of Previous Readmissions. These are variables that I believe will improve the performance of my model. This addition yielded 56 variables model. The most significant variable is the Medicaid Insurance with a p-value of 0.043. The second statistically significant variable is Systolic Blood Pressure at admission with a p-value of 0.06. The addition of these variables has slightly improved the C-Statistic score to 0.65.

Finally, adding Echo data variable has slightly improved the C-Statistic score to 0.66 and did not change the predictable variables from the previous model except with the addition of the number of previous readmissions variable. Although there was no big improvement in prediction accuracy compared to the Adhere model variables, I intend to use this model for the remaining of the study as Echo data variable might show statistical significance when characterizing very early readmissions. Table 4.6 below provide detail information about the comparison of the three models.

Characteristics of the Study Cohort	
Characteristics	**Study Cohort (Visits=5488)**
Demographics	
Mean Age	60.32
Male	3675 (66%)
Female	1875 (34%)
Mean Admission Weight, Kg	91.5
Mean Discharge Weight, Kg	90
Mean Change Weight, Kg	-2
White	3394 (61%)
Black, African American	1354 (24%)
Primary Insurance	
Medicare	3334 (60%)
Medicaid	1299 (23%)
Commercial	685 (12%)
Medical History	
Coronary artery disease	95 (2%)
Prior myocardial infarction	192 (3%)
Atrial fibrillation	239 (4%)
Congestion	29 (<1%)
Chronic obstructive pulmonary disease	145 (2.5%)
Chronic renal insufficiency	896 (16%)
Diabetes	233 (4%)
Fatigue	32 (<1%)
Hyperlipidemia	9 (<1%)
Hypertension	177 (3%)
Peripheral edema	16 (<1%)
Peripheral vascular disease	8

Rales	2
Stroke/transient ischemic attack	39 (<1%)
Ventricular tachycardia/ventricular fibrillation	481 (8.6%)
Laboratory values	
Mean Admission B-type natriuretic peptide, pg/ml	1220.68
Mean Discharge B-type natriuretic peptide, pg/ml	944.12
Mean Admission Blood urea nitrogen,	36.06
Mean Discharge Blood urea nitrogen,	34.21
Mean Admission Creatinine, mg/dL	2.17
Mean Discharge Creatinine, mg/dL	1.94
Mean Admission Hemoglobin, g/dL	11.09
Mean Discharge Hemoglobin, g/dL	10.47
Mean LVEF, %	41.55
Mean Admission NACL, mEq/L	134.83
Mean Discharge NaCl, mEq/L	135.23
In hospital Medications	
Diuretics	3529 (63.5%)
Statins	473 (8.5%)
Beta Blockers	423 (7.6%)
ACE	419 (7.5%)
Digoxin	140 (2.5%)
Nitrates	205 (3.7%)
Aldosterone Receptor Antagonist	271 (4.9%)
Angiotensin receptor blocker	90 (1.6%)
Echo Data	
Normal Echo Test	92 (1.6%)
Abnormal Echo Test	666 (12%)
Unaccessible Echo Test	4197 (75.6%)

Vital Signs	
Mean Admission Systolic blood pressure, mm Hg	125.23
Mean Discharge Systolic blood pressure, mm Hg	117.45
Mean Admission Heart rate, beats/min	87.75
Mean Discharge Heart rate, beats/min	80.46
Mean Length of Stay, Days	7.64
Mean Number of Previous Readmissions, visits	5.74

Table 4.5: Characteristics of the Study Cohort

30 Days Readmission Models	Best Predictable Variables	P-Value	C-Statistic
Adhere Variables	SBP at Admission Medicaid	0.0138 0.06	0.64
Adhere Variables + Medication + Vitals at Discharge + Length of Stay + Number of Previous Readmissions	Medicaid SBP at Admission	0.0438 0.06	0.65
Adding Echo Data	Medicaid SBP at Admission Number of Previous Readmissions	0.0358 0.0417 0.0232	0.66

Table 4.6: Models Comparison

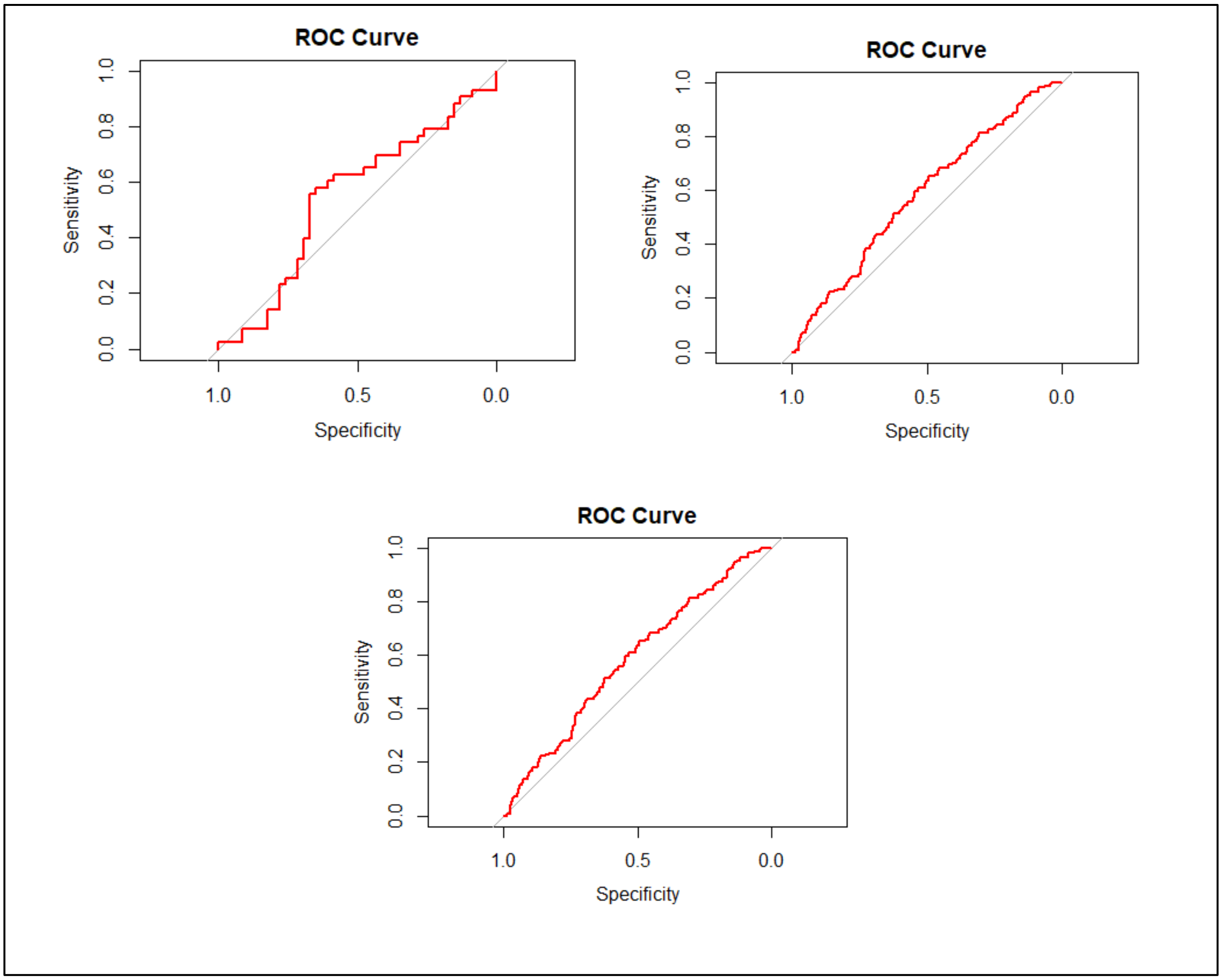


Figure 4.6: ROC Curves for the three Models

After I chose the third model, I started running 6-days readmission models to find the cut off when the patient characteristics started to change. Results of the logistic regression models among the six days ranges show that the variables Diuretics, Angiotensin Receptor Blocker, and Abnormal Echo test are predictable variables for 1-6, and 2-7 days range of readmission. Table 4.7 contains detailed visit characteristics for the six days ranges. After the seven days readmission, the visit characteristics start to change at readmission at 3-8 days. 770 visits of the 2030 30-readmission visits occur in the first seven days (~%40) — the ranges 3-8, 4-9, 5-10,6-11, and 7-12 days of until readmission have no similarity in their visit characteristics. Then, we start to see the similarity in visit characteristics for the ranges 8-13, 9-14, and 9-15 days until readmission. The common predictable variables for these ranges are the number of previous readmissions, and diabetes. After the 9-14 readmission range, I switched my ranges to be seven days instead of six, and this is due to the lower number of visits for the ranges of 10 days readmission and further. (See figure 4.5) The third group of visits happens in the ranges of 16-27 days readmissions.

Readmission Range of Days	**Best Predictable Variables**	**P-Value**	**C-Statistic**
1-6 Days	Creatinine at Admission **Diuretics** **Angiotensin Receptor Blocker** **Abnormal Echo**	0.0102 0.0243 0. 0189 0.0366	0.72
2-7 Days	Male Medicaid **Diuretics** **Angiotensin Receptor Blocker** **Abnormal Echo**	0. 0427 0. 0163 0.0274 0.0457 0.05	0.73
3-8 Days	Medicaid NACL at Discharge	0.0220 0.0365	0.68
4-9 Days	NACL at Discharge	0.05	0.73
5-10 Days	Heart Rate at Discharge Number of Previous Readmissions	0.02732 0.02009	0.68
6-11 Days	BNP at Admission Heart Rate at Discharge Number of Previous Readmission	0.0156 0.0149 0.0242	0.73

7-12 Days	BNP at Admission Number of Previous Readmissions	0.002405 0.000516	0.78
8-13 Days	Male **BNP at Admission** **Diabetes** **Number of Previous Readmission**	0.034058 0.024930 0.006265 0.000565	0.76
9-14 Days	Weight Change Value Commercial Insurance **BNP at Admission** **Diabetes** **Number of Previous Readmission** Age	0.0144 0.034590 0.002948 0.002669 0.001095 0.028851	0.76
9-15 Days	Heart Rate at Discharge **Diabetes** **Number of Previous Readmissions**	0.0452 0.0305 0.0049	0.67
10-16 Days	Medicare	0.0292	0.61
11-18 Days	Number of Previous Readmissions ACE	0.0027 0.078	0.66
12-19 Days	ACE Number of Previous Readmissions	0.0184 0.00074	0.71
13-20 Days	Medicaid Medicare ACE Number of Previous Readmissions	0.0466 0.0295 0.05 0.07	0.65
14-21 Days	History of Myocardial Infarction Age	0.0156 0.05	0.70
15-22 Days	Age Weight Change Value	0.009 0.0183	0.72

	Medicare Systolic Blood Pressure at Admission History of Myocardial Infraction **Angiotensin Receptor Blocker** **Number of Previous Readmissions**	0.0338 0.0165 0.0102 0.023 0.0222	
16-23 **Days**	BUN at Admission Creatinine at Admission COPD **Angiotensin Receptor Blocker** **Number of Previous Readmissions**	0.05 0.05 0.0352 0.0114 0.061	0.69
17-24 **Days**	BUN at Admission Creatinine at Admission COPD **Angiotensin Receptor Blocker** **Number of Previous Readmissions**	0.016 0.0429 0.083 0.0061 0.0277	0.64
18-25 **Days**	Male BUN at Admission Commercial Insurance **Diuretics** **Angiotensin Receptor Blocker** Aldosterone Receptor Antagonist	0.058 0.06 0.0215 0.054 0.021 0.0416	0.65
19-26 **Days**	Male BUN at Admission Commercial Insurance Creatinine at Admission **Diuretics** **Angiotensin Receptor Blocker** Aldosterone Receptor Antagonist	0.0224 0.0220 0.0493 0.0366 0.0360 0.0128 0.0359	0.72
20-27 **Days**	Male Commercial Insurance **Diuretics**	0.08 0.0262 0.0335	0.72

	Statins	0.0189	
	Angiotensin Receptor Blocker	0.0027	
	Aldosterone Receptor Antagonist	0.0076	
21-28 Days	Male	0.057	0.79
	Hispanic	0.0436	
	Medicaid	0.0019	
	Medicare	0.00128	
	Other Insurance	0.0032	
	BUN at Admission	0.0213	
	Creatinine at Admission	0.0098	
22-29 Days	Beta Blockers	0.0416	0.64
	Diuretics	0.069	
23-30 Days	Male	0.0151	0.77
	Medicaid	0.0309	
	Medicare	0.0390	
	Creatinine at Admission	0.061	
	Heart Rate at Admission	0.0353	
	Statins	0.0147	
	Aldosterone Receptor Antagonist	0.0420	

Table 4.7: Visit Characteristics in different windows for readmission

After knowing the clusters of the 30 days readmission, I ran three logistic models. For the first cluster (1-7 days), Angiotensin Receptor Blocker is the most statistically significant variable with a p-value of 0.0222. The second most statistically significant variable is Diuretics with a p-value of 0.0225. The third most statistically significant variable is Abnormal Echo with a p-value of 0.0251. The logistic regression model C-Statistic score for this model is 0.72. The second cluster is 8-15 days where the number of previous readmissions is the most statistically significant variable with a p-value of 0.0019. The second predictable variable is Diabetes with a p-value of 0.0372. The accuracy of the prediction model is the same as the first cluster at 0.63.

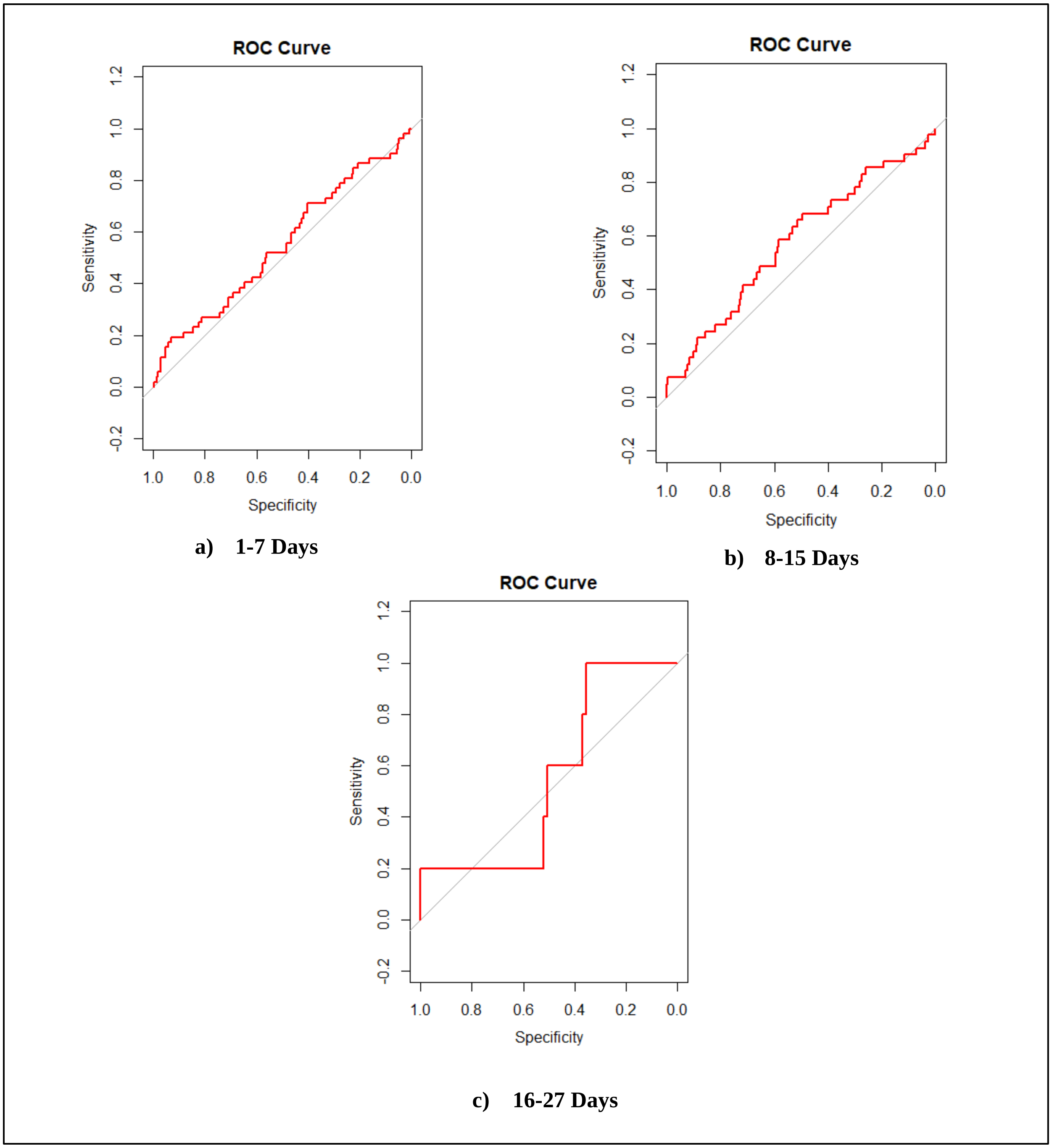
Finally, the last cluster is readmissions from 16-27 days. This has a slightly higher C-Statistic score with 0.71. The two statistically significant variables are Creatinine at admission and History of Myocardial Infarction with a p-value of 0.0214 and 0.0383 sequentially. Table 4.8 shows a detailed list of the statistically significant variables with their p-values. Figure 4.7 shows the ROC curve for the three clusters.

Readmission Range of Days	**Best Predictable Variables**	**P-Value**	**C-Statistic**
1-7 Days	Medicaid	0.0367	0.72
	NACL at Discharge	0.0314	
	Diuretics	0.0225	
	Angiotensin Receptor Blocker	0.0222	
	Abnormal ECHO	0.0251	
8-15 Days	Diabetes	0.0372	0.63
	Number of Previous Readmissions	0.0019	
16-27 Days	Creatinine at Admission	0.0214	0.71
	History of Myocardial Infarction	0.0383	
	Number of Previous Readmissions	0.0489	

Table 4.8: Visits Characteristics of the Three Clusters

Variable	**Value**
PPV	0.89
NPV	0.25
Sensitivity	0.99
Specificity	0.02

Table 4.9: Very Early Readmission Model Performance Measures



a) 1-7 Days

b) 8-15 Days

c) 16-27 Days

Figure 4.7: ROC for the Three Clusters

4.4 DISCUSSION

The current models that exist in the literature have overlooked shorter than 30 days of readmission. For the small number of studies that have looked into shorter than 30 days of readmission, they have not scientifically shown why they choose the number that they investigated (mostly seven days). It is essential to try clustering the 30 days readmission and define the very early readmission groups to improve the discharge decision process. I have added new variables in my model when predicting 30-readmission that have not been applied in other models such as the number of previous readmissions and Echo data. I believed they would improve the prediction accuracy, yet, the model prediction accuracy was within what currently exists in the literature with C-Statistic of 0.66.

This study does not aim to provide a model for predicting 30-days readmission as this field have been extensively investigated. The study brings a new look to the field that has been mostly overlooked, which is defining when very early readmission occurs. In this study, I was able to define very early readmission to be readmission within the first seven days. These visits share similar characteristics. Also, they count for ~40% of the 30-day readmissions. I hypothesize that "very early re-admission" are those readmissions that are more likely due to error, or readmissions, that if avoided, could have led to substantial cost savings and/or potentially improved health outcomes. Knowing the characteristics of the very early readmission at the time of discharge will help the healthcare provider to prevent their occurrence. I have used a logistic regression model for predicting very early readmissions where its C-Statistic score is 0.72. The model has a positive predictive value of 0.89 and a negative predictive value of 0.25. (see table 4.9)

Diuretics, Angiotensin Receptor Blocker, and Abnormal Echo test were consistent with the ranges of 1-6, and 2-7 days of readmission. Also, it is statistically significant when running the model for 1-7 days of readmission. The three most predictable for very early readmission are Diuretics, Angiotensin Receptor Blocker, and Abnormal Echo test. For the second cluster, from 8-15 days, it does not have an overlapping variable with the very early readmission cluster. This could support the clusters assumption where no overlapping variable exists. The two most predictable variables in the second cluster are Diabetes and number of previous readmissions. The second clusters have a lower C-Statistic score compared to the other cluster at 0.63. Finally, the third cluster, 16-27 days, have a better C-Statistic at 0.71. The two most predictable variable are Creatinine at admission and History of Myocardial Infarction and the number of previous readmissions where the later variable considered a statically significant variable for the first cluster.

In this study, I have used a logistic regression model to run my analysis. Logistic regression algorithm has widely been used in the health predictive model literature for many reasons. First, it is easy to interpret, especially for a non-machine learning expert. Also, it is a well-behaved algorithm that can be trained as long the problem can be linearly separable. However, some cons of using logistic regression include and not limited to; it does not handle a large number of categorical features. Also, it tends to have high bias and low variance. This might affect the model accuracy. For that, I will be investigating different machine learning algorithm in the next chapter to see if that will change my result and improve my model accuracy.

Also, the current dataset suffers from data missingness, especially in the lab values. (Figure 4.4) The missingness of data in Brain Natriuretic Peptide (BNP) at discharge reached 82%. The BNP at admission variable is missing 36% of the time. Missingness of data could as well affect the performance of the model. Applying Imputation technique to solve the missingness of the data could have a positive impact and improve the model prediction accuracy. In the next chapter, I will test if applying a different machine learning algorithm will change “the very early readmission” definition and if the three clusters with their most statistically significant variables will be different. Finally, I will be applying the imputation technique to see if that will improve the model accuracy or not.

Chapter 5. IMPROVING ON THE MODEL

The goal of this chapter is to improve on the very early readmission prediction model that I presented in chapter 4. In the previous chapter, using the software R and the package caret, I separated the data into training and testing sets using ten folds cross-validation to minimize the mean squared error (MSE). Then, I build a Logistics regression model from the training data using the generalized linear model function in R **glm()**. The outcome of the model is binary (readmitted or not). The prediction accuracy of the model built for 30-day readmission is measured by the C-Statistic score, which was 0.66. C-Statistic is a measure of goodness of fit for binary outcomes in a logistic regression. (Stephanie, 2016) I have built a five days range prediction models to characterize day 1-30 readmissions (shorter than 30 days). In that effort, I concluded that readmission within eight days shares similar characteristics that are different from the rest of the 30-day readmissions. The C-Statistic score for day 1-7 prediction model is 0.72. I hypothesize that this group of visits are the very early readmission which is more likely due to error, or readmissions, that if avoided, could have led to substantial cost savings and/or potentially improved health outcomes.

In this chapter, I report on my attempts to improve the prediction accuracy by applying different machine algorithm. There has been a plethora of research in building readmission prediction models. When building a prediction model in healthcare, researchers favored using logistic regression algorithm. (Yang et al., 2016) (Ross et al., 2008) It has been widely used in medical and biomedical research mainly to formulate models that determine which factors help determine whether an outcome happens. Logistic Regression is a great tool for binary classification. The output of logistic regression is more informative than other classification algorithms since it expresses the relationship between an outcome variable and the features.(Murphy, 2012) (Yang et al., 2016) On the other hand, there are other machine learning algorithms that researchers have used to solve classification problems in the healthcare field such as predicting readmission. Such algorithms include but are not limited to Random Forest, Support Vector Machine, Deep Unified Network (DUN), and Regularized Logistic regression (LASSO). In the literature, there is no single machine learning algorithm proved to have better prediction accuracy. (Garcia-Arce et al., 2017)(Golas et al., 2018)(Yang et al., 2016) In my study, I selected Random Forest and LASSO algorithms to test if that will improve the prediction accuracy for predicting 30 days readmission. Then, I chose the algorithm with the higher prediction accuracy for 30 days readmission to conduct my windowing days. I want to see if the model will result in different clusters for visits within 30-days readmissions and if the definition for the very early readmissions will be the same as the one from the previous chapter.

Finally, I test if applying Imputation technique to solve the missingness of the data will improve the model accuracy. Missing data could influence the accuracy of any prediction model. Electronic Health Record (EHR) data missingness is a result of not having been explicitly collected for research purposes. Missing data could happen for various reasons such as lack of collection, or lack of documentation. (Wells et al., 2013) In the EHR, many “NULL” values are assumed to be negative, which makes mitigating missing value difficult. (Wells et al., 2013) The missingness of data could affect our understanding of patient care, specifically in readmissions.

5.1 APPLYING DIFFERENT MACHINE LEARNING ALGORITHMS

In this chapter, I replicate the work I have done on chapter 4 but with different machine learning algorithms to see if this will lead to a better prediction accuracy for 30 days readmission and different clusters for 30-day readmissions. I want to compare the result from logistic regression with other non-linear models such as Random Forest. Random Forest is a tree-based classification method that can capture nonlinearity. In random forests, an ensemble learning method for classification, a large number of binary tree classifiers are trained separately and then combined into a single prediction. Each classifier is a decision tree where the tree "votes" for that class and the forest selects the classification that has the most vote. Random forest algorithm can handle thousands of input variables without variable deletion. Also, it estimates missing data and maintains accuracy when large number of data is missing. (Breiman, 1999)

When building the random forest model, I used the same set up from the previous chapter. I build the model using R Software, applying the library (randomForest). I split the data into training and testing sets using 10 folds cross-validation. In the random forest, a class label is represented by the leaf at the bottom, and the internal node represents the decision that needs to be made based on features. I started training the random forest model of 171 trees each of them on p/3 predictors. Similar to the logistic regression model, I use 58 variables. The more trees in the model the more complex of the random forest model. Also, the node size is another critical factor in the random forest model is the node size. The node size represents the minimum sample size of terminal nodes, the larger the sample size in terminal nodes, the less complex of the tress, the less complex of the random forest model. The number of trees selected, and node size is tuned each time I fit a random forest using 10-fold cross-validation. I chose the model with minimum prediction error. First, using the trained model, I predict on the testing data. Then, I get the true outcome values for the testing data. Then, I measured the accuracy of the model using (caret) package and applied **confusionMatrix()** function which summarizes the prediction performance of a classification model. This function reports metrics such as Accuracy, Sensitivity, and Specificity.

Then, I chose Regularized Logistic regression (LASSO), an algorithm known for its robustness in dealing with high dimensional data, avoiding overfitting the data, and providing high accuracy. This makes it usually the default method in many supervised machine learning tasks. (Yang et al., 2016) Moreover, it is considered state-of-the-art in readmission prediction tasks. (Futoma et al., 2015; Yang et al., 2016).

5.2 APPLYING IMPUTATION METHOD

Dealing with missing data is a universal problem across different domains when it comes to prediction analysis. Medical and Biomedical research is not an outlier when it comes to dealing with missing data. Although the availability and accessibility of EHR data advance the research of patient-centered outcome, the missing data in EHR could affect the validity of any research conclusion. In EHR missing data, it is difficult to separate between missing data and negative value. (Wells et al., 2013) Figure 5.1 below from Wells et al. 2013 gives an overview of the missing data issue in EHR systems with several options to solve this issue.

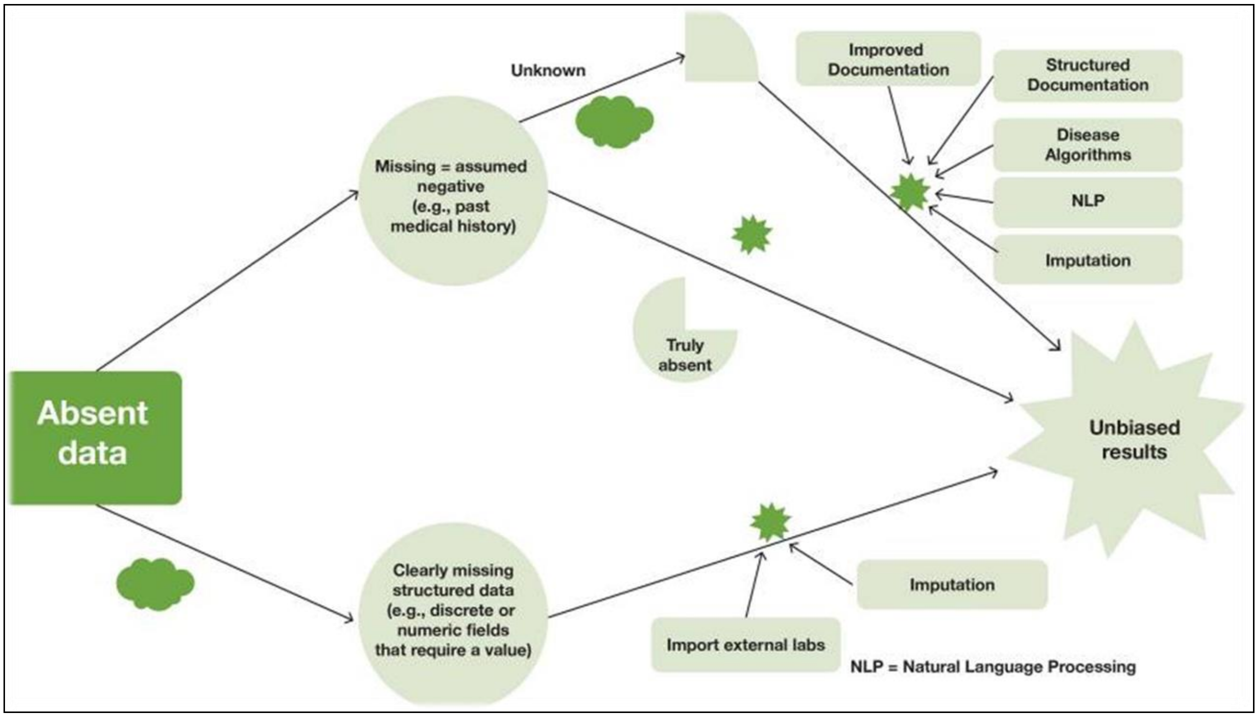

Unknown
Improved Documentation
Structured Documentation
Disease Algorithms
NLP
Imputation
Missing = assumed negative (e.g., past medical history)
Truly absent
Absent data
Unbiased results
Clearly missing structured data (e.g., discrete or numeric fields that require a value)
Imputation
Import external labs
NLP = Natural Language Processing

Figure 5.1: Overview of the missing data problem with electronic health records [Wells et al. 2013]

When Imputing that data, we need to understand what kind of missing data we are dealing with. Missing data falls into three categories; missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). MCAR occurs when the probability that the data point is missing is not related to the value of that data point or any other variable. On the other hand, MAR occurs when the probability that the data point is missing depends on the known values but not the value of the missing data. MNAR is the most extreme situation which occurs when the probability that the data point is missing depends only on the value of that data point or another unmeasured variable(s). (Wells et al., 2013)

My dataset suffers from missing data for the different lab tests and vital signs. Figure 5.2 shows a histogram for the missing variable where Brain natriuretic peptide (BNP) test at discharge is the most missing variable with (82%). In my analysis and giving that I am dealing with lab and vital sign tests, I chose to implement the extreme case where I assumed that the missing data are not at random. When dealing with retrospective data, it is hard to assume MAR without contacting the patients directly. (Wells et al., 2013)

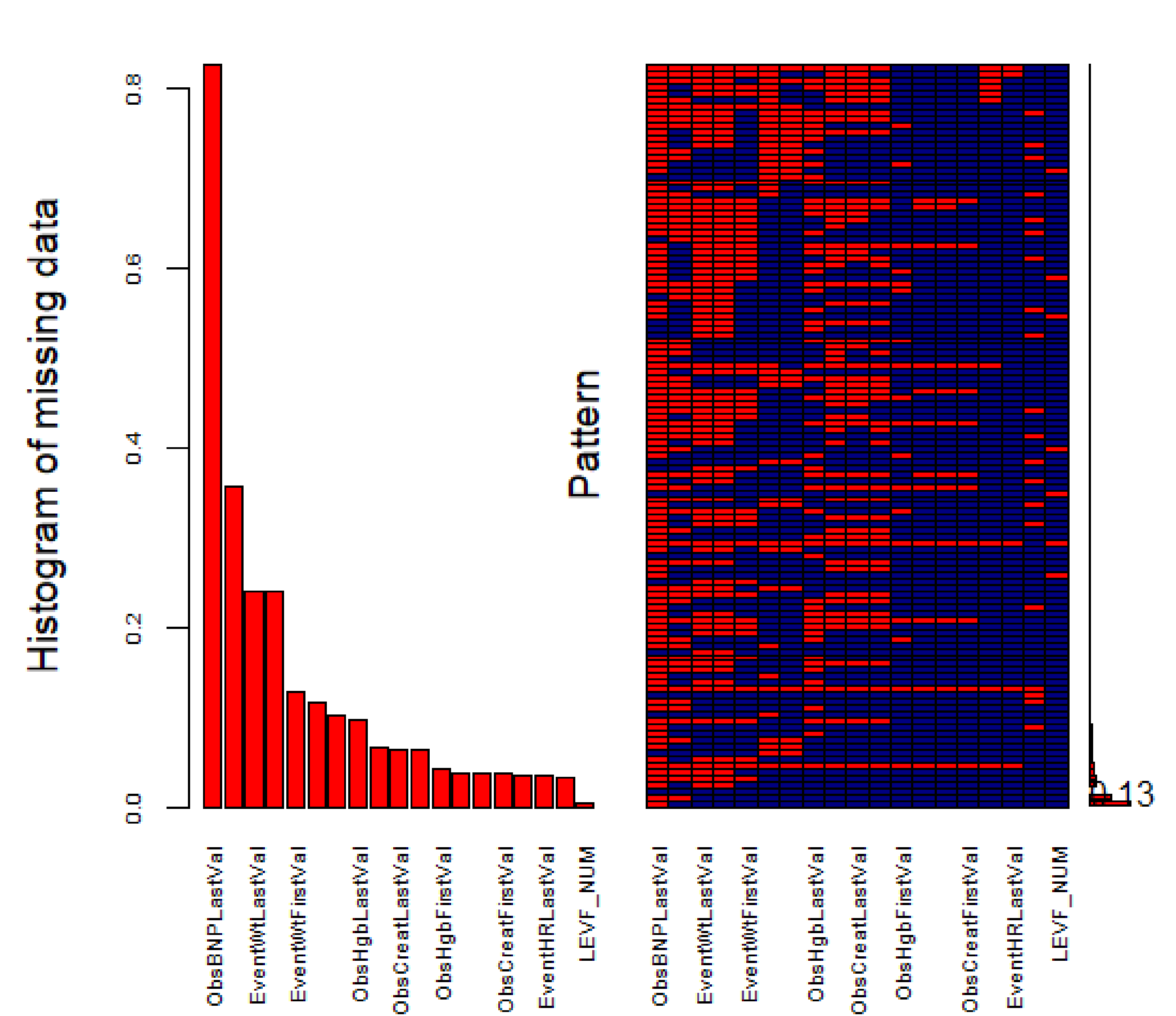


Figure 5.2: Histogram of missing data

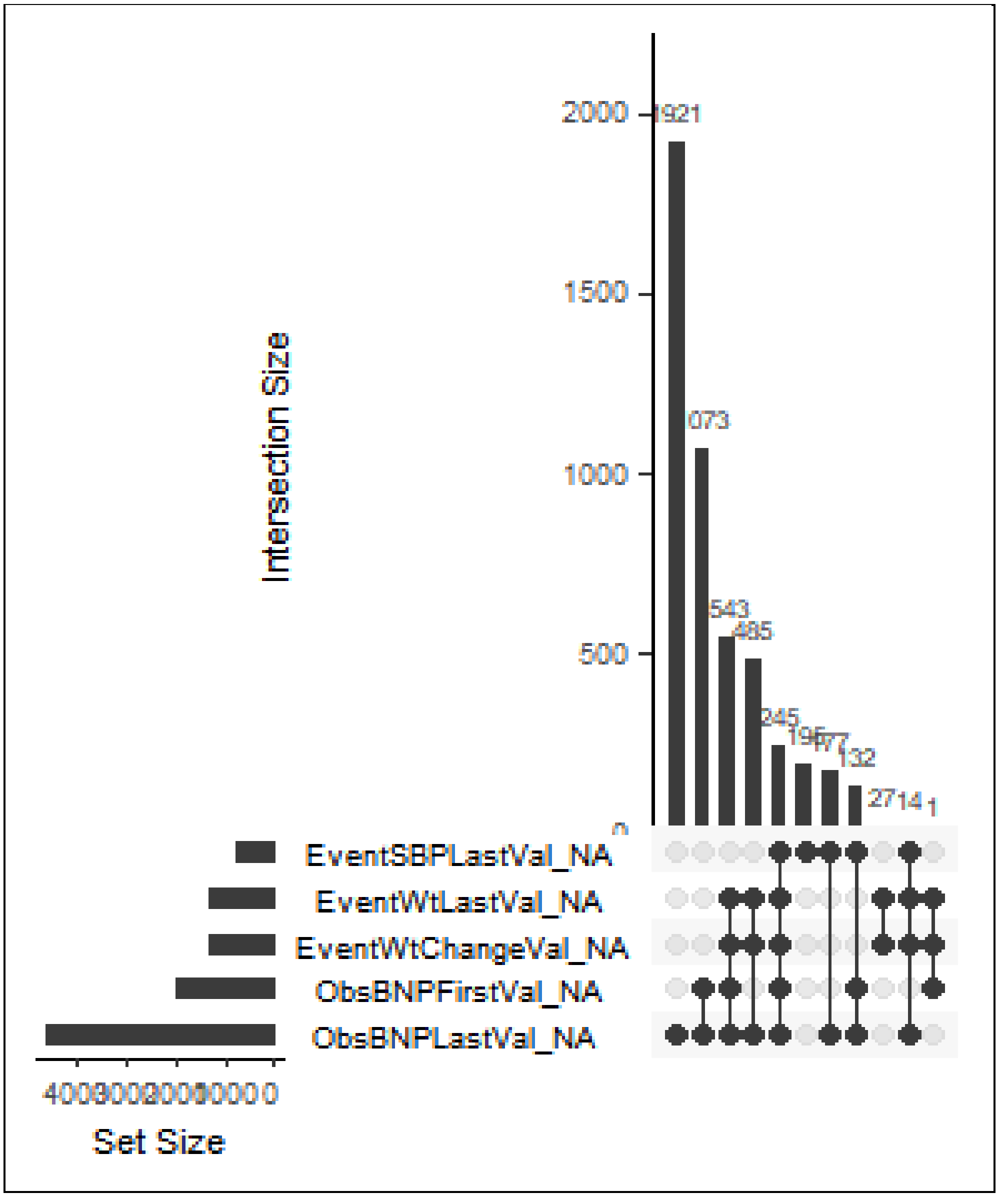


Figure 5.3: Intersection of Missing Variables among variables

5.3 RESULTS

I have started my analysis using the same cohort from the previous chapter of 5488 visits and 1123 patients. When running my random forest model, I have used the same 56 variables from the previous chapter. Of those variables evaluated for the risk of 30 readmissions, Creatinine at admission, Systolic Blood Pressure at admission, Blood Urea Nitrogen (BUN) at discharge, and

Diuretics are the most significant variables. The accuracy of the random forest model to predict 30 readmission has a 0.68 C-Statistic score with 0.88 sensitivity and 0.22 specificity. In my study, the random forest algorithm provides a slightly improved prediction score for 30-day readmission than applying logistic regression algorithm which is 0.66.

In the random forest, three essential factors affect the model; the number of trees, node size, and the number of variables randomly sampled as candidates at each split. Giving the problem I am solving here is a logistic problem, the number of variables randomly sampled is p/3 where p is the number of variables. In order to select the optimal number of trees and the node size, I have fit a random forest using 10-fold cross-validation and tuned these two variables each time. I chose the model with minimum prediction error which comes to 280 trees with node size of 20. Random forest algorithm uses Mean Decrease Gini, which gives ranking scores of the variables, the larger the score, the more importance of the model. The mean decrease in Gini coefficient is a measure of how each variable contributes to the homogeneity of the nodes and leaves in the resulting random forest. (Menze et al., 2009) Table 5.1 below shows the most important variables ordered by its importance top down.

Variables	Mean Decrease Gini
Creatinine at Admission	9.90
SBP at Admission	9.63
BUN at Discharge	9.53
Diuretics	1.1
Beta Blockers	0.91
BNP at Admission	7.96
BUN at Admission	7.91

Heart Rate at Admission	7.90
Number of Previous Readmissions	7.86
Abnormal Echo	0.8
Last Weight Value	7.23
History of VT	0.70
Hemoglobin at admission	7.06
LEVF Number	7.05

Table 5.1: List of Important variables for Random Forest Model

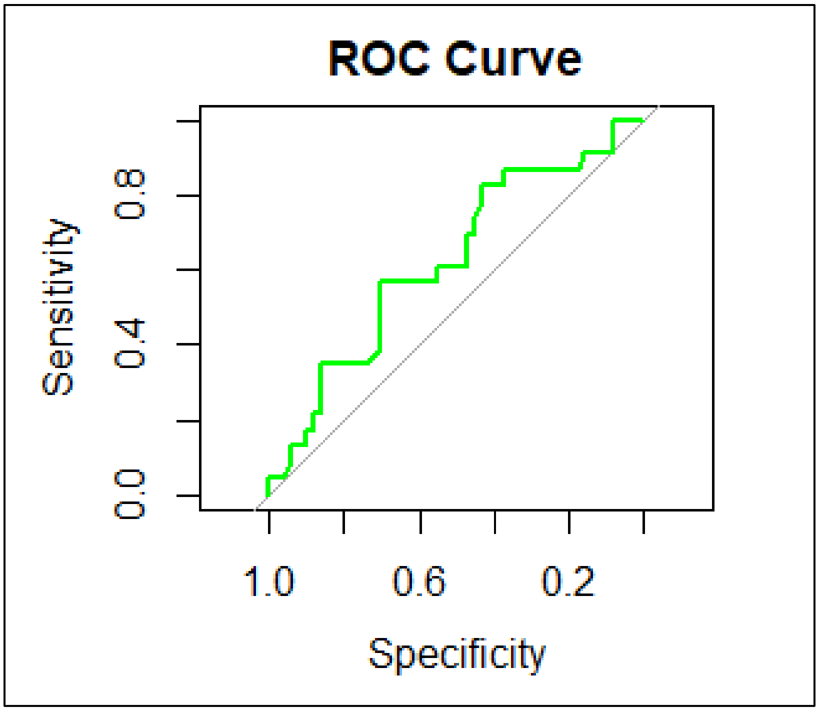


Figure 5.4: ROC Curve for Random Forest

The imputation does improve the accuracy of the random forest but by a small margin. The accuracy of the random forest prediction model after applying the imputation is 0.69 with 0.95 sensitivity and 0.13 specificity. The number of trees with a minimum prediction error is 280 trees with node size of 20. After using Mean Decrease Gini, which ranks the variables important to the model, the three most important variables are; Hemoglobin at Admission, Systolic Blood, Brain natriuretic peptide (BNP) at discharge, and Weight Change value. Table 5.2 below shows a detailed list of the most important variables.

Variables	**Mean Decrease Gini**
Hemoglobin at Admission	64.24
BNP at Discharge	63.71
Weight Change Value	61.9
LEVF_NUM	61.12
Number of Previous Readmissions	60.64
SBP at Discharge	60.52
History of Ventricular tachycardia/ Ventricular fibrillation	5.67
Caucasian	5.37
Statins	5.07
Age	59.32
BNP at Admission	58.25
Last Weight Value	58.06
First Weight Value	57.93
Hemoglobin at Discharge	56.28
Systolic Blood Pressure at Admission	55.53

Table 5.2: List of Important variables for Random Forest Model (After Imputation)

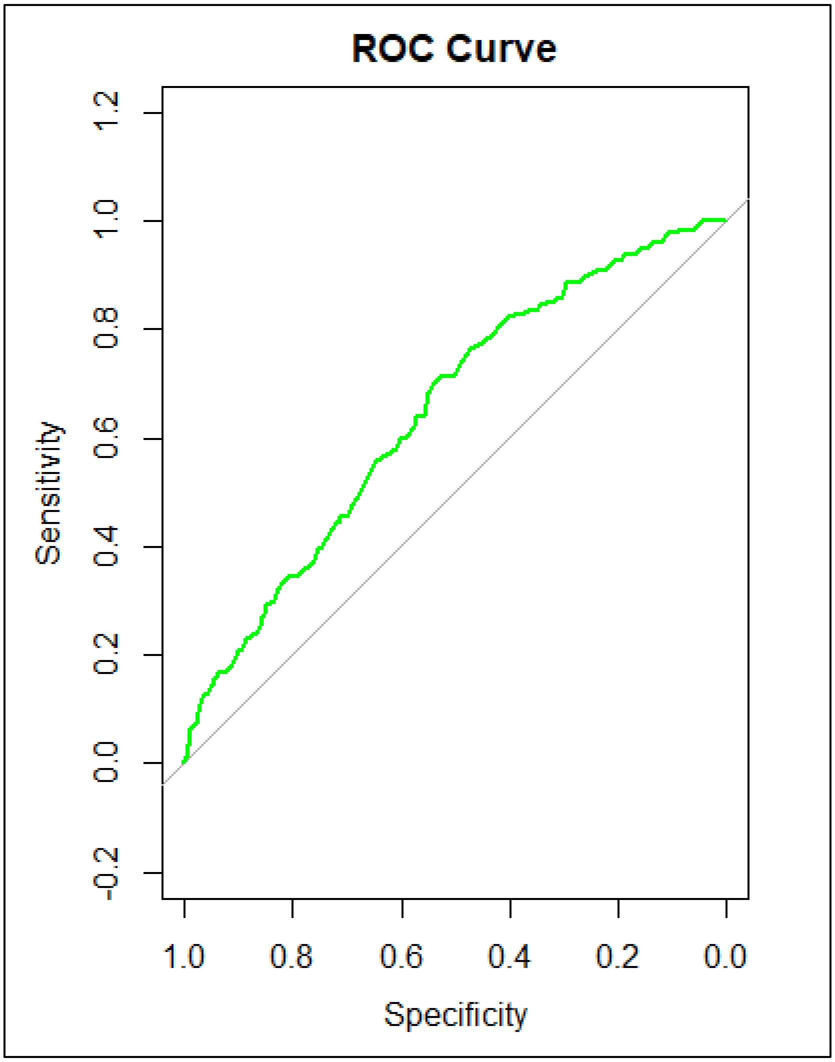


Figure 5.5: ROC Curve for Random Forest (with imputation)

Similar to Random Forest, LASSO did not provide higher prediction accuracy than Logistic regression for 30 days readmission. LASSO require fully observed data since it cannot handle missing data. For that, I used the complete dataset from section 5.3 where the imputation assumed to be missing not at Random (MNAR). LASSO C-Statistic score for 30 days readmission was 0.63. Of the 58 variables evaluated for the risk of 30 readmissions, the number of previous readmissions is the most predictable variable with a p-value of 7.70e-12. The second most significant variable is Hemoglobin at admission with a p-value of 1.97e-05. Finally, the third significant variable is Age with a p-value of 0.0002. Figure 5.7 below is a plot of the lasso model using cross-validation. The lowest point in the plot corresponds to the best model.

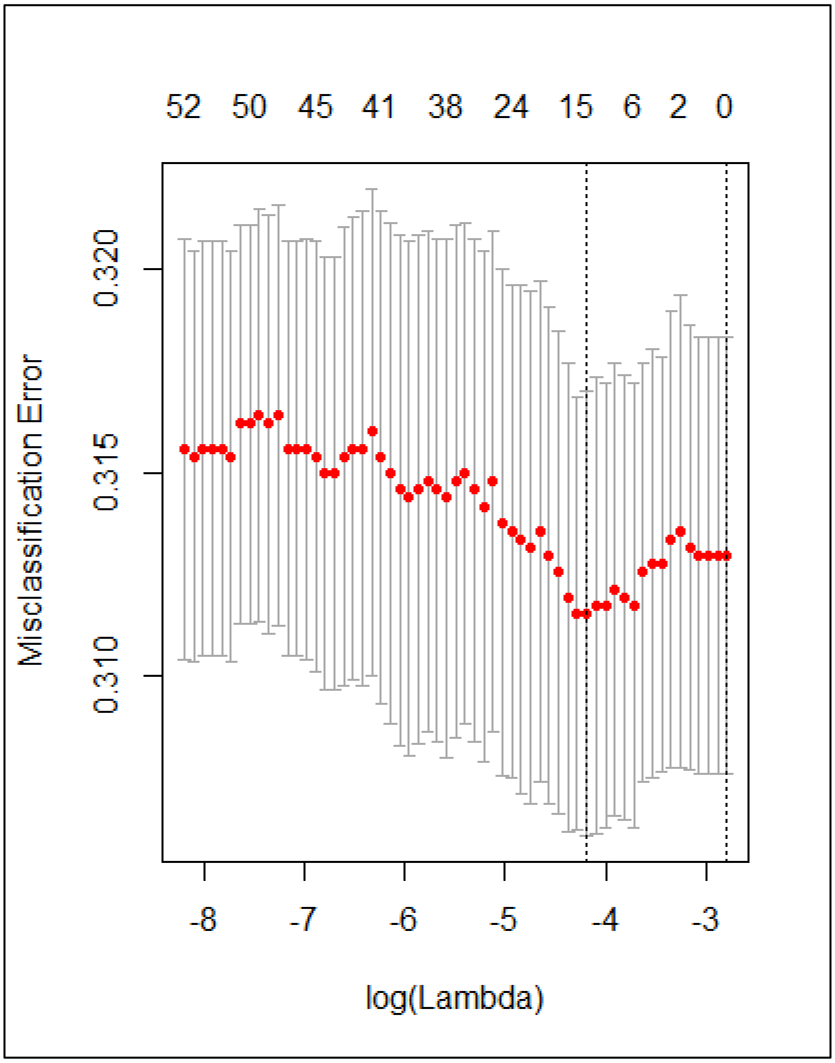


Figure 5.6: Plot of the Lasso model using cross-validation

Figure 5.8 below shows the path trajectory of the fitted sparse regression parameters. Each curve in the figure shows how the regression coefficient of variable changes according to the value of lambda. It should be read from right to left (lambda from small to large). The variables stay in the last to become zero means that they are probably significant as we need to impose a large lambda to make them zero. On the other hand, the variables which quickly become zero are probably weak or insignificant variables.

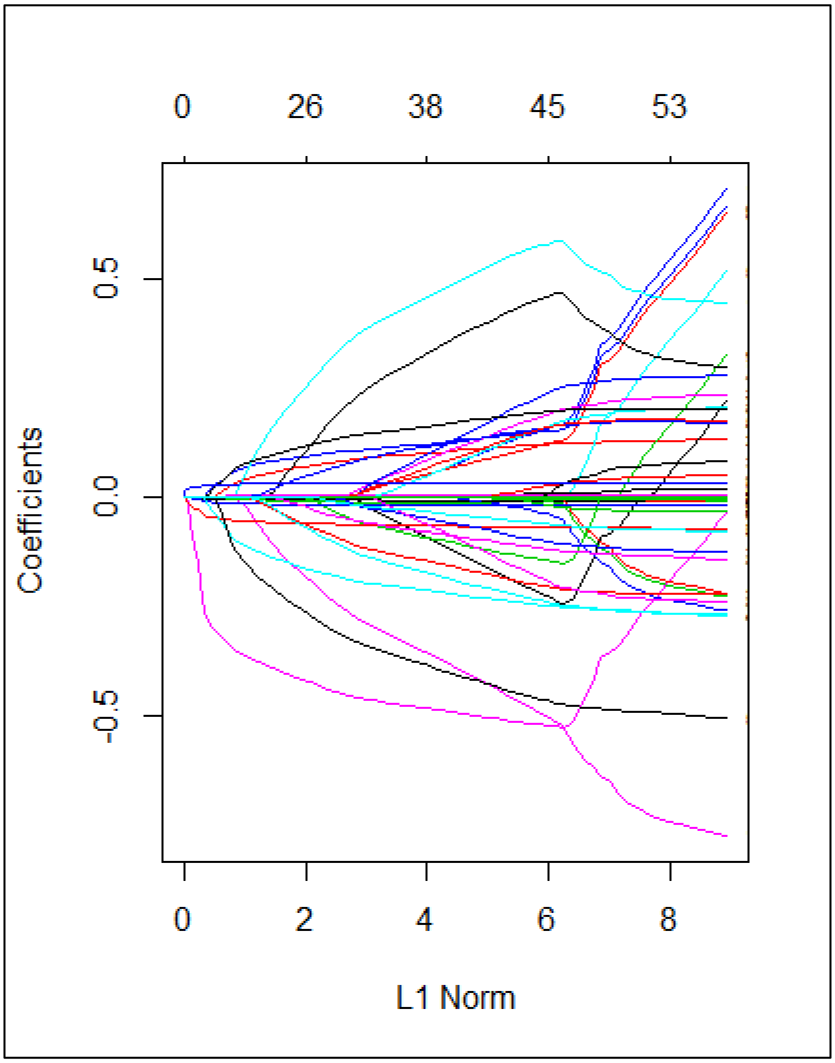


Figure 5.7: Path Trajectory of the fitted sparse regression parameters

Despite it did not improve prediction accuracy compared to Logistic regression, I wanted to see if applying the Random Forest algorithm will yield different clusters for the 30 days readmission. Also, if the definition for very early readmission is continued to be within seven days of readmission. I chose Random Forest with Imputation because it yielded the highest accuracy when predicting 30 readmissions. Running the model on a seven days range did not yield any clusters. The variable that repeatedly showed to be significant in the different ranges is BNP at discharge. Table 5.3 below shows detailed information about the different models such as the significance variables, and the model accuracy for each model.

Readmission Range of Days	**Best Predictable Variables**	**Model Accuracy**
1-6 Days	**BNP at Discharge** Heart Rate at Admission Creatinine at Admission SBP at Admission First Weight Value	0.91
2-7 Days	BNP at Admission **BNP at Discharge** Weight Change Value Creatinine at Discharge Creatinine at Admission SBP at Admission	0.88
3-8 Days	LEVF Number BNP at Admission Weight Change Value Hemoglobin at Discharge BUN At Discharge **BNP At Discharge**	0.88
4-9 Days	LEVF Number BNP at Admission Weight Change Value SBP at Discharge	0.89

	BNP At Discharge BUN at Discharge	
5-10 **Days**	LEVF Number Weight Change Value **BNP At Discharge** BUN at Discharge SBP at Discharge Last Weight Value	0.91

Table 5.3: Visit Characteristics on seven days ranges, Random Forest with Imputation

5.4 DISCUSSION

This study does not aim to provide a model for predicting 30-days readmission as this field have been extensively investigated. Instead, my research brings a new look to the field that has been mostly overlooked, which is defining when very early readmission occurs. In this study, applying different machine learning algorithm have not improved my prediction accuracy for 30-days readmission. Neither Random Forest nor LASSO provided great improvement in prediction accuracy than logistic regressions. Compared to logistic regression C-Statistic score of 0.66, Random Forest has slightly higher accuracy, and LASSO yielded lower accuracy with C-Statistic score of 0.68 (0.88 sensitivity and 0.22 specificity) and 0.63 respectively. The number of previous readmissions is the significant variable that was consistent among the different machine learning algorithms with and without applying imputation for the 30 days of readmission. (See table 5.5)

Also, the current dataset suffers from data missingness, especially in the lab values. (Figure 5.2) The missingness of data in Brain Natriuretic Peptide (BNP) at discharge reached 82%. The BNP at admission variable is missing 36% of the time. Missingness of data could as well affect the performance of the model. Applying Imputation technique to solve the missingness of the data could have a positive impact and improve the model prediction accuracy. However, the imputation with the assumption that the missing is not at random did not improve the model accuracy for Logistic regression to predict 30 days readmission. The C-Statistic score dropped from 0.66 to 0.64 with Imputation. Also, in a random forest, there was no significant improvement in prediction accuracy for 30 days readmission as the C-Statistic score improved from 0.68 to 0.69.

Moreover, applying the imputation technique to build the LASSO model did not result in a better model than logistic regression with no imputation. The C-Statistic score for LASSO model is 0.63. The characterization for the 30 days readmissions did not yield different clusters for the Random Forest model. The imputation could be a reason that affected such changes. I believe it is an area that could be improved. This requires consulting a specialist to understand the clinical procedures of ordering lab tests for heart failure patients. Which tests score depends on the ordering of which lab tests? In statistical language, it is to understand the interaction effect among lab test. The most predictive variables for the different ranges models in Random Forest is Brain Natriuretic Peptide (BNP) at discharge. The accuracy of the different models was around 0.89.

30 Days Readmission Models	**Best Predictable Variables**	**P-Value**	**C-Statistic**
Logistic Regression Without Imputation	Medicaid	0.0358	0.66
	SBP at Admission	0.0417	
	Number of Previous Readmissions	0.0232	
Logistic Regression WITH Imputation	Age	0.00043	0.64
	Male	0.0407	
	Hemoglobin at Admission	2.27e-06	
	Creatinine at Admission	0.0227	

	Creatinine at Discharge	0.0433	
	NACL at Discharge	0.0253	
	Number of Previous Readmissions	4.88e-11	
Random Forest Without Imputation	Creatinine at Admission		0.68
	SBP at Admission		
	BUN at Discharge		
	Diuretics		
	Beta Blockers		
	BNP at Admission		
	Number of Previous Readmissions		
	Abnormal Echo		
Random Forest WITH Imputation	Hemoglobin at Admission		0.69
	BNP at Discharge		
	LEVF		
	Number of Previous Readmissions		
	SBP at Discharge		
	Caucasian		
	History of Ventricular tachycardia/ Ventricular fibrillation		
LASSO WITH Imputation	Age	0.00026	0.63
	Caucasian	0.0165	
	Asian	0.0011	
	Hemoglobin at Admission	1.97e-05	
	NACL at Discharge	0.0030	
	Heart Rate at Discharge	0.0064	
	Beta Blockers	0.0081	
	Angiotensin Receptor Blocker	0.0279	
	Number of Previous Readmissions	7.70e-12	

Table 5.4: Algorithms Comparison for 30 days readmission model

Chapter 6. CONCLUSION

In this chapter, I start by summarizing my research journey, identify my research contribution and its broader implication for both researchers and clinicians, and discuss my research limitations. Finally, I will explore future directions for my research beyond the scope of this dissertation.

6.1 RESEARCH SUMMARY

In this dissertation, I have described my work using retrospective EHR data to identify subgroups for readmission within 30 days for heart failure patients (i.e., shorter than 30 days). I have demonstrated that there exist different groups at risk for readmission within 30 days and one of these groups are the very early readmission group.

In Chapter 2, I provide the background of prediction models for heart failure readmission, understanding the past and the present of the field. More specifically, I investigated prior studies that studied the risk of readmission for heart failure patients and studies that identified heart failure patient characteristics measured before discharge, where the prediction model was presented in the study. I conducted an updated systematic review searching the literature for studies that investigated shorter than 30 days readmission.

In Chapter 3, I built an accurate visit table that accurately captures temporal information about all admission and discharges for heart failure patients. I used retrospective 260,776 EHR data points from UW Medical Health Systems ranges from 2010 to 2017 to build the visit table. The accuracy of the table was then validated via UW Medicine Annual Report, Discharge Summary Report, Patient Chart Review, and consulting a cardiologist.

In Chapter 4, using the visit table created from chapter 3, I characterized day 1-30 readmissions and defined the range of days within 30 days readmission where patients shared similar characteristics. In this process, I built prediction models for six days ranges to capture when patient group characteristics started to change. This characterization yielded three groups, which are day 1-7, 8-15, 16-27. In this chapter, I was able to define very early readmission to occur in the first seven days. Also, I highlight the best predictor variables for very early readmission.

In Chapter 5, I tried to improve upon the prediction model from chapter 4. In my approach, I applied different machine learning algorithms and applied an imputation technique for missing data to see if this will lead to improved model accuracy and/or change the definition of very early readmission.

6.2 BROADER IMPLICATION

I believe that the work that I have accomplished in this dissertation will have a broader implication for both the research and the clinical setting. Although the focus of this research is on heart failure patients, the same methodology used in creating the visit table and characterizing 30 days readmission could be applied to different diseases. The border implication of this research is that it leverages the high volume of clinical data and highlights the existence of different groups with different characteristics within 30 days readmission. If applying the same methodology on different diseases, this could potentially improve the discharge decision making and hence improve the patient care and lower the treatment cost.

6.2.1 *Research Implication*

For researchers, the work and the code I built to clean the data and create the visit table in chapter 3 shows how the process is surprisingly difficult. As part of the process, the researchers need to assess the visit table quality via manual chart review and patient discharge summary report. The process of assessing the visits table quality is time-consuming as it requires clinical expertise.

In future work, I would like to understand which of my steps and methods for creating the visit table might generalize to other institutions. To the extent that they generalize, I could build a tool that would help researchers automate the process of building an appropriate visit table and reduce the need for time-consuming chart review. This tool would produce a “cleaned” visit table, which could then be the nucleus for a number of analyses, including prediction tasks.

6.2.2 *Clinical Implication*

For Clinicians, identifying and flagging heart failure patients with risk of very early readmission at the time of discharge could serve as a powerful clinical decision support tool. My idea of using retrospective data to understand the relationship between the patient characteristics at the time of discharge and the duration of their next readmission will potentially improve the discharge decision process. Being able to identify and flag patients with risk very early readmission, which is 40% from the 30 days readmission visits, could potentially lower the treatment cost and the risk of mortality for heart failure patients.

Embedding these prediction models into a good effective decision support tool in the clinical setting could unleash new discharge protocols for the group of patients who are at risk of very early readmission. For example, if knowing (at the discharge time) that the patient falls in this category, and after the case review of two specialists, the discharge protocol could be that the patient will only be discharged if they have a caregiver. In Chapter 4, I showed how my model that will flag patient with risk of very early readmission have an accuracy of 0.89. With my UW population characteristics, my model has a Positive predictive value of 0.89 and a negative predictive value of 0.25. The model has a sensitivity of 0.99 and specificity of 0.02.

6.3 RESEARCH LIMITATIONS

My research showed different groups that exist within 30 days readmission and defined very early readmission to be within the first seven days, using EHR from UW Medicine Health System. However, it has limitations that could be addressed to improve the result and make it generalizable to different health systems. In general, my dissertation sheds light on an overlooked yet important group of readmission patients that are ~40% of the entire group of 30 days readmission patients. However, it was not tested on a different health institute EHR, used mostly structured EHR data, need improvement on the imputation technique used, and no direct access to the identified EHR.

Perhaps the most significant limitation of my dissertation is working only on one data source (one institute) that is the UW Medicine health system. The result from chapter 4 and the definition of very early readmission along with the characteristics associated with it cannot be generalized since the model was not tested on a different EHR system. Using one EHR health system from one geographic location could make the result specific only to the UW Health System population. Table 4.5 from chapter 4 about the characteristics of the study cohort shows that the majority (61%) are from the white race.

Furthermore, not having direct access to the EHR and using only the EHR data that was queried and hashed the patient ID, visit id and the patient date of birth as described in Chapter 3. This can be seen in the missing of some important variables that exist in the identified EHR. The patient age is an example of a variable that has 17% data missing. Although I was able to fix this and reduce it to 3%, having direct access to the identified EHR will eliminate such missingness for important variables. Also, having access to identified EHR will allow tracking a specific patient to see if the reason for not being readmitted is healthy or that they have died via The Social Security Death Index. Currently, a death that occurs outside the hospital is not well documented, and only death that happens inside the hospital is.

Also, missing not at random imputation technique that used in chapter 5 need to be improved since the assumption made about the missingness was not based on domain expert. The need to understand the ordering procedure for lab tests for heart failure patient is essential to understand the dependency among the test variables. For example, if test A is >50 then skip test B, otherwise, conduct test B. In this example, it is clear that test B depends on the value of A.

Finally, using mostly structured data could limit the conclusiveness of the analysis. Including free text such as physician notes, will bring valuable insight and a new dimension to my analysis. There are critical unstructured data that resides in the EHR if included in the analysis it will strengthen the dissertation hypothesis and might improve the model accuracy.

6.4 FUTURE DIRECTIONS

This dissertation has several potential rooms of improvement and area of expansion for both the research and clinical world. In my effort to build an accurate visit table in chapter 3, I discovered how creating an accurate visit table is a difficult process and requires time-consuming manual work to assess its quality. This shows a need in building a tool that would help researchers automate the process of building an appropriate visit table and reduce the need for time-consuming chart review and discharge summary reports. This tool would produce a “cleaned” visit table, which could then be the nucleus for a number of analyses, including prediction tasks.

Also, in Chapter 3, I described the data source of my analysis which is UW medicine health systems. To increase the validity and generalizability of my work, I plan on testing my model on a different health system data from a different geographic location.

In Chapter 4, I discussed the variables that I used in building the model. While I included new variables that were not used in the literature such as Echocardiogram Test data, which shows its significance in predicting very early readmission, free text data was mostly overlooked. Including the free text data that either resides in the EHR, patient chart review and discharge summary report could improve the accuracy and validity of my findings. This will require new skill for my analysis that is natural language processing (NLP). Moreover, in Chapter 4, the model shows the most significant variables in terms of their p-value. To make this human readable for the clinical use, I will provide the exact range that is considered significant for the variables. This information will be more valuable to the clinicians in real time use than just flagging the patient.

The overall goal of this research is to embed these prediction models into a good effective decision support tool in the clinical setting to improve on the discharge decision process. The tool would assess the patient characteristics at the time of discharge. The tool will flag patient that might be at risk of having very early readmission and highlight the most significant variables (such as Echo Test and Sodium Chloride Test). The flag on the patient will require the sign off from some specialists. The cardiologists then could have a second review of the patient chart and discharge summary report and look into the highlighted most significant variables and decide on the course of action. The tool will not make a decision; it will only flag patients at risk of very early readmission and provide that information to the specialist.

6.5 Final Conclusion

In this dissertation, I have shown my work on using retrospective electronic health record data to characterize 1-30 days readmission for heart failure patients. I conducted a literature review to understand the field of predictive analytics for heart failure patients, explicitly searching the definition of very early readmission that is shorter than 30 days. This work showed how shorter than 30 days readmission received little research attention in the literature. This yielded my research questions; can I identify subgroups for readmission within 30 days (i.e., shorter than 30 days)?

In the process of answering my research questions, I have built an accurate visit built and discovered that this process was surprisingly difficult. The process of creating the visit table showed the poor information that currently exists in the EHR about what count as a readmission due to billing procedure and human error. This visit table is the nucleus for building my prediction model when characterizing day 1-30 readmission for heart failure patient. My dissertation showed the existence of three subgroups within the 30 readmissions that share the same visit characteristics; day 1-7, 8-15, and 16-27. Moreover, I was able to define very early readmission to be within day 1-7. These visits account for 40% of the 30 days readmission. This group of patients also have an average of 6 readmissions. Given that heart failure is the leading cause for readmission, the single largest expense for Medicare in the last 16 years and has an overall five years mortality rate of 60%; identifying very early readmission at the time of discharge could be a valuable clinical tool.

BIBLIOGRAPHY

American Heart Association. (2013). What is Echocardiography. Retrieved from www.heart.org/idc/groups/heart.public/@wcm/@hcm/documents

Beaulieu-Jones, B. K., Lavage, D. R., Snyder, J. W., Moore, J. H., Pendergrass, S. A., & Bauer, C. R. (2018). Characterizing and Managing Missing Structured Data in Electronic Health Records: Data Analysis. *JMIR Medical Informatics*, *6*(1), e11. https://doi.org/10.2196/medinform.8960

Benjamin, E. J., Virani, S. S., Callaway, C. W., Chamberlain, A. M., Chang, A. R., Cheng, S., … Muntner, P. (2018, March 20). Heart Disease and Stroke Statistics—2018 Update: A Report From the American Heart Association. https://doi.org/10.1161/CIR.0000000000000558

Breiman, L. (1999). Random Forests, 5–32. Retrieved from http://machinelearning202.pbworks.com/w/file/fetch/60606349/breiman_randomforests.pdf

Clarke, A. (1990). Are readmissions avoidable? *BMJ (Clinical Research Ed.)*, *301*(6761), 1136–1138. https://doi.org/10.1136/bmj.301.6761.1136

Donzé, J., Lipsitz, S., David, B., & Jeffrey, S. (2013). Causes and patterns of readmissions in patients with common comorbidities: Retrospective cohort study. *BMJ (Online)*, *347*(December), 1–12. https://doi.org/10.1136/bmj.f7171

Eastwood, C. A., Howlett, J. G., King-Shier, K. M., McAlister, F. A., Ezekowitz, J. A., & Quan, H. (2014). Determinants of early readmission after hospitalization for heart failure. *Canadian Journal of Cardiology*, *30*(6), 612–618. https://doi.org/10.1016/j.cjca.2014.02.017

Ejection Fraction Heart Failure Measurement | American Heart Association. (n.d.). Retrieved March 21, 2019, from https://www.heart.org/en/health-topics/heart-failure/diagnosing-heart-failure/ejection-fraction-heart-failure-measurement

Fonarow, G. C. (2003). The Acute Decompensated Heart Failure National Registry (ADHERE): opportunities to improve care of patients hospitalized with acute decompensated heart failure. *Reviews in Cardiovascular Medicine*, *4 Suppl 7*, S21-30. https://doi.org/10.1007/s13398-014-0173-7.2

Fonarow, G. C., Abraham, W. T., Albert, N. M., Gattis, W. A., Gheorghiade, M., Greenberg, B., … Young, J. (2004). Organized program to initiate lifesaving treatment in hospitalized patients with heart failure (OPTIMIZE-HF): Rationale and design. *American Heart Journal*, *148*(1), 43–51. https://doi.org/10.1016/j.ahj.2004.03.004

Futoma, J., Morris, J., & Lucas, J. (2015). A comparison of models for predicting early hospital readmissions. *Journal of Biomedical Informatics*, *56*, 229–238. https://doi.org/10.1016/j.jbi.2015.05.016

Gabayan, G., Asch, S., Hsia, R., Zingmond, D., Liang, L.-J., Han, W., … PhDg, Robert E. Weiss, PhDg, and Benjamin C. Sun, MD, M. (2013). Factors Associated with Short-Term Bounce-back Admissions Following Emergency Department Discharge. *Annals of Emergency Medicine*, *62*(2), 136–144. https://doi.org/10.1016/j.annemergmed.2013.01.017.Factors

Garcia-Arce, A., Rico, F., & Zayas-Castro, J. L. (2017). Comparison of Machine Learning Algorithms for the Prediction of Preventable Hospital Readmissions. *Journal for Healthcare Quality*, *40*(3), 1. https://doi.org/10.1097/JHQ.0000000000000080

Goff Jr, D. C., Pandey, D. K., Chan, F. A., Ortiz, C., & Nichaman, M. Z. (2000). Congestive Heart Failure in the United States: Is There More Than Meets the I(CD Code)? The Corpus Christi Heart Project. *Archives of Internal Medicine*, *160*(2), 197–202. https://doi.org/10.1001/archinte.160.2.197

Golas, S. B., Shibahara, T., Agboola, S., Otaki, H., Sato, J., Nakae, T., … Jethwani, K. (2018). A machine learning model to predict the risk of 30-day readmissions in patients with heart failure: A retrospective analysis of electronic medical records data. *BMC Medical Informatics and Decision Making*, *18*(1), 1–17. https://doi.org/10.1186/s12911-018-0620-z

Hersh, A., Masoudi, F., & Allen, L. (2013). Postdischarge environment following heart failure hospitalization: expanding the view of hospital readmission. *Journal of the American Heart Association*, *2*(2), 1–14. https://doi.org/10.1161/JAHA.113.000116

Hospital Readmissions Reduction Program (HRRP). (2018). Retrieved December 26, 2018, from https://www.cms.gov/medicare/medicare-fee-for-service-payment/acuteinpatientpps/readmissions-reduction-program.html

Kheirbek, R. E., Fletcher, R. D., Bakitas, M. A., Fonarow, G. C., Parvataneni, S., Bearden, D., … Ahmed, A. (2015). Discharge Hospice Referral and Lower 30-Day All-Cause Readmission in Medicare Beneficiaries Hospitalized for Heart Failure. *Circulation. Heart Failure*, *8*(4), 733–740. https://doi.org/10.1161/CIRCHEARTFAILURE.115.002153

Kociol, R. D., Shaw, L. K., Fonarow, G. C., O’Connor, C. M., Hernandez, A. F., Reyes, E. M., … Felker, G. M. (2011). Admission, Discharge, or Change in B-Type Natriuretic Peptide and Long-Term Outcomes. *Circulation: Heart Failure*, *4*(5), 628–636. https://doi.org/10.1161/circheartfailure.111.962290

Lund, L., Rich, M., & Hauptman, P. (2018). Complexities of the Global Heart Failure Epidemic. *Journal of Cardiac Failure*, *24*(12), 813–814. https://doi.org/10.1016/j.cardfail.2018.11.010

Menze, B. H., Kelm, B. M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., & Hamprecht, F. A. (2009). A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics*, *10*, 1–16. https://doi.org/10.1186/1471-2105-10-213

Miñana, G., José, M., Núñez, E., Mollar, A., Santas, E., Valero, E., … Núñez, J. (2017). European Journal of Internal Medicine Length of stay and risk of very early readmission in acute heart failure. *European Journal of Internal Medicine*, *42*, 61–66. https://doi.org/10.1016/j.ejim.2017.04.003

Murphy, K. (2012). *Machine Learning, A Probabilistic Perspective*.

O’Connor, M., Murtaugh, C. M., Shah, S., Barrón-Vaya, Y., Bowles, K. H., Peng, T. R., … Feldman, P. H. (2016). Patient Characteristics Predicting Readmission among Individuals Hospitalized for Heart Failure. *Medical Care Research and Review*, *73*(1), 3–40. https://doi.org/10.1177/1077558715595156

Ross, J. S., Mulvey, G. K., Stauffer, B., Patlolla, V., Bernheim, S. M., Keenan, P. S., & Krumholz, H. M. (2008). Statistical models and patient predictors of readmission for heart failure: a systematic review. *Archives of Internal Medicine*, *168*(13), 1371–1386. https://doi.org/10.1001/archinte.168.13.1371

Rutten, F. H., Clark, A. L., & Hoes, A. W. (2016). How big a problem is heart failure with a normal ejection fraction? *BMJ (Clinical Research Ed.)*, *353*, i1706. https://doi.org/10.1136/bmj.i1706

RxNorm. (n.d.). Retrieved March 20, 2019, from https://www.nlm.nih.gov/research/umls/rxnorm/

Stephanie. (2016). C-Statistic: Definition, Examples, Weighting and Significance - Statistics How To. Retrieved December 28, 2018, from https://www.statisticshowto.datasciencecentral.com/c-statistic/

US Census Bureau. (n.d.). Retrieved March 20, 2019, from https://www.census.gov/en.html

Voigt, J., Sasha, J., Taylor, A., Krucoff, M., Reynolds, M., & Gibson, M. (2014). A reevaluation of the costs of heart failure and its implications for allocation of health resources in the united states. *Clinical Cardiology*, *37*(5), 312–321. https://doi.org/10.1002/clc.22260

Wells, B. J., Nowacki, A. S., Chagin, K., & Kattan, M. W. (2013). Strategies for Handling Missing Data in Electronic Health Record Derived Data. *EGEMs (Generating Evidence & Methods to Improve Patient Outcomes)*, *1*(3), 7. https://doi.org/10.13063/2327-9214.1035

Yang, C., Delcher, C., Shenkman, E., & Ranka, S. (2016). Predicting 30-day all-cause readmissions from hospital inpatient discharge data. *2016 IEEE 18th International Conference on E-Health Networking, Applications and Services, Healthcom 2016*, 1–6. https://doi.org/10.1109/HealthCom.2016.7749452

VITA

Ahmad Khalid Aljadaan was born in Riyadh, Saudi Arabia and studied at King Fahd University of Petroleum and Minerals (KFUPM) where he earned a bachelor’s degree in Management Information Systems. He went on to pursue a graduate degree in Information Science at the University of Michigan, Ann Arbor and earned a master’s degree there before moving to Stanford, CA to work as an academic researcher. While working at Stanford University, he developed an interest in data science and decision analysis in healthcare and decided to pursue his interests. He earned his Doctoral of Philosophy in Biomedical and Health Informatics from the University of Washington in 2019.