A STRATEGY FOR REDUCING CONGESTIVE HEART FAILURE READMISSIONS THROUGH THE USE OF INTERVENTIONS TARGETED BY MACHINE LEARNING PREDICTION

A Dissertation

Presented to

The Graduate Faculty at the University of Akron

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

James Natale

May, 2015
A STRATEGY FOR REDUCING CONGESTIVE HEART FAILURE READMISSIONS THROUGH THE USE OF INTERVENTIONS TARGETED BY MACHINE LEARNING PREDICTION

James Natale

Dissertation

Approved:

Advisor
Dr. Shengyong Wang

Committee Member
Dr. Chien-Chung Chan

Committee Member
Dr. Richard Einsporn

Committee Member
Dr. Ping Yi

Committee Member
Dr. Jiang Zhe

Accepted:

Department Chair
Dr. Sergio Felicelli

Dean of the College
Dr. George K. Haritos

Interim Dean of the Graduate School
Dr. Rex D. Ramsier

Date
ABSTRACT

Hospitals are faced with a financial penalty when patients are readmitted to the hospital within thirty days of discharge. However preventing readmissions is a difficulty for hospitals. There are a wide variety of interventions which are aimed towards improved patient outcomes, as well as preventing patient readmissions. However these interventions are expensive in terms of financial outlay, as well as in the time spent by staff. However, if interventions are applied indiscriminately, these methods would cost more than they would have saved due to readmissions prevention. We describe an all encompassing strategy which would allow for hospitals to reduce their readmissions in a cost-effective manner. We apply an analytical approach to all aspects of the problem. By creating a predictive model with machine learning methods on hospital records, we can determine the risks of patients being readmitted. We detail how the literature on intervention strategies can be condensed and utilized to determine prospective strategies which may be of interest. Utilizing the risk predicted by the model, as well as published literature on interventions we determine the optimal solution of which patients will receive which interventions through a genetic algorithm heuristic search. Only by combining the three aspects together can we formulate an analytics driven approach to reducing readmissions in a cost-effective manner.
# TABLE OF CONTENTS

| LIST OF TABLES | vii |
| LIST OF FIGURES | ix |

## CHAPTERS

### I. INTRODUCTION

1.1. Research Motivation .......................................................... 1
1.2. Research Objectives .......................................................... 2
1.3. Dissertation Overview .......................................................... 2

### II. RESEARCH BACKGROUND AND LITERATURE REVIEW

2.1. Rise of Chronic Disease and the Readmission Crisis ...................... 4
2.2. Problem Justification ............................................................ 8
   2.2.1. Optimization Problem .................................................. 8
2.3. Literature Review .................................................................. 10
   2.3.1. Reducing Readmissions ................................................ 10
   2.3.2. Predicting Readmission Risk .......................................... 11

### III. RISK PREDICTION

3.1. Data Collection .................................................................... 14
3.2. Data Cleaning ...................................................................... 17
3.3. Data Mining .......................................................................... 18
   3.3.1. Initial Investigation ....................................................... 21
3.3.2. Feature Selection........................................................................................................27
3.3.3. Feature Selection Discussion with Domain Experts.................................33
3.3.4. Cost Sensitive Learning..........................................................................................34
3.3.5. Selection of Modeling Approach.................................................................36
3.3.6. Calculating Patient Risk......................................................................................40
3.3.7. Finalization of Modeling Approach.................................................................41
  3.3.7.1. The Class Imbalance Problem.................................................................41
  3.3.7.2. Learning Algorithm.....................................................................................41
  3.3.7.3. Feature Selection.........................................................................................42

IV. OBTAINING INTERVENTION OPTIONS FROM LITERATURE.........................43
  4.1. Efficacy..................................................................................................................46
  4.2. Methodology Followed..........................................................................................49
    4.2.1. Example 1.......................................................................................................50
    4.2.2. Example 2.......................................................................................................51
    4.2.3. Example 3.......................................................................................................52
  4.3. Summarized Literature on CHF Readmission Interventions ......................54
  4.4. Formulation of Estimates for Experimentation.................................................55
  4.5. Discussion of Results............................................................................................57

V. OPTIMIZATION OF IMPLEMENTATION THROUGH HUERISTICS ..........59
  5.1. Understanding the Genetic Algorithm............................................................65
  5.2. Enterprise Economic Concerns..........................................................66
  5.3. Results of Experimentation................................................................................69
    5.3.1. Base-Line Measures.....................................................................................69
5.3.2. Results Interpretation........................................................................70
5.4. Sensitivity Analysis..............................................................................72
5.5. Comparison to Other Solution Heuristics.............................................84

VI. FUTURE RESEARCH..............................................................................86
6.1. Method for Applying Approach in Practice .........................................86
6.2. Intervention Options............................................................................87
6.3. Sampling..............................................................................................88
6.4. Missing Data Imputation......................................................................90
   6.4.1. Drop Features...............................................................................90
   6.4.2. Drop Patients...............................................................................91
   6.4.3. Imputation....................................................................................91
6.5. Portability of Research Framework to Other Hospitals......................93
6.6. A Question of Root Cause....................................................................96
6.7. Differentiation Between Necessary and Unnecessary Readmissions ....98
6.8. Summary of Results............................................................................98

BIBLIOGRAPHY..........................................................................................100
<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15</td>
</tr>
<tr>
<td>2</td>
<td>20</td>
</tr>
<tr>
<td>3</td>
<td>28</td>
</tr>
<tr>
<td>4</td>
<td>28</td>
</tr>
<tr>
<td>5</td>
<td>29</td>
</tr>
<tr>
<td>6</td>
<td>29</td>
</tr>
<tr>
<td>7</td>
<td>30</td>
</tr>
<tr>
<td>8</td>
<td>30</td>
</tr>
<tr>
<td>9</td>
<td>31</td>
</tr>
<tr>
<td>10</td>
<td>44</td>
</tr>
<tr>
<td>11</td>
<td>47</td>
</tr>
<tr>
<td>12</td>
<td>49</td>
</tr>
<tr>
<td>13</td>
<td>50</td>
</tr>
<tr>
<td>14</td>
<td>51</td>
</tr>
<tr>
<td>15</td>
<td>53</td>
</tr>
<tr>
<td>16</td>
<td>54</td>
</tr>
</tbody>
</table>
17 Finalized List of Eight Programs Considered for Heuristics…………………..55
18 Finalized List of Eight Programs Considered for Heuristics…………………..56
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Example Receiver Operator Characteristic curve</td>
<td>20</td>
</tr>
<tr>
<td>2</td>
<td>Example of the Rapid Miner Process creation</td>
<td>22</td>
</tr>
<tr>
<td>3</td>
<td>Representation of the k-fold cross-validation which is often performed</td>
<td>22</td>
</tr>
<tr>
<td>4</td>
<td>Rapid Miner process for predicting CHF readmissions through comparing ROC curves</td>
<td>23</td>
</tr>
<tr>
<td>5</td>
<td>Sub-process of the compare ROC's operation, which gives the modeling algorithms and necessary transforms</td>
<td>23</td>
</tr>
<tr>
<td>6</td>
<td>Initial performance in readmission prediction</td>
<td>24</td>
</tr>
<tr>
<td>7</td>
<td>Entirety of a random forest tree for CHF readmission prediction simply guesses 'no' for every patient</td>
<td>25</td>
</tr>
<tr>
<td>8</td>
<td>Representation of the SMOTE algorithm creating synthetic points</td>
<td>26</td>
</tr>
<tr>
<td>9</td>
<td>ROC curve obtained through data post feature selection and oversampling with an optimistic bias</td>
<td>35</td>
</tr>
<tr>
<td>10</td>
<td>ROC curve obtained through data post feature selection and oversampling with a pessimistic bias</td>
<td>36</td>
</tr>
<tr>
<td>11</td>
<td>ROC curve for the prediction of CHF patient readmissions under univariate feature selection</td>
<td>38</td>
</tr>
<tr>
<td>12</td>
<td>ROC Curve for patient readmissions under the features selected by the tree-based feature selection</td>
<td>38</td>
</tr>
<tr>
<td>13</td>
<td>Performance of the random forest on the classification of CHF readmissions with SMOTE</td>
<td>39</td>
</tr>
<tr>
<td>14</td>
<td>Performance of the random forest on the classification of CHF readmissions without SMOTE</td>
<td>39</td>
</tr>
</tbody>
</table>
15 Final leaf nodes of an example patient ................................. 40
16 Genetic algorithm which was implemented to determine the optimal solution for applied interventions. .................................................. 60
17 Representation of the two point crossover mechanism which was utilized........ 61
18 Patient risk is laid out sequentially, with patients at highest risk first. .................. 61
19 Representation of the solution vector format used .................................. 62
20 A view of the population characteristics as generations increase .................. 72
21 Showing model performance evaluation as the generation increases.............. 74
22 A plot of the difference in the fitness values as the generation increases......... 75
23 Plot which shows the starting point and the length for the vector of application of program 1 to patients. ........................................... 76
24 Plot which shows the starting point and the length for the vector of application of program 2 to patients. ........................................... 76
25 Plot which shows the starting point and the length for the vector of application of program 3 to patients. ........................................... 76
26 Plot which shows the starting point and the length for the vector of application of program 4 to patients. ........................................... 77
27 Plot which shows the starting point and the length for the vector of application of program 5 to patients. ........................................... 77
28 Plot which shows the starting point and the length for the vector of application of program 6 to patients. ........................................... 77
29 Which shows the starting point and the length for the vector of application of program 7 to patients .................................................. 78
30 Which shows the starting point and the length for the vector of application of program 8 to patients .................................................. 78
31 The performance under the 'inflated' rate ........................................... 81
32 The performance under the best estimated rate ......................................... 82
33 The performance under the 'deflated' rate ........................................... 83
Performance comparisons of several machine learning algorithms and the changes in performance due to sample size ................................................................. 89
CHAPTER I

INTRODUCTION

1.1. Research Motivation

The purpose of this research is the formulation of a holistic strategy for hospital systems to follow in order to financially optimize their efforts to reduce hospital readmissions. For delivering health care, health care organizations are ultimately concerned with the readmissions of patients. The largest diagnosis group which impact hospital readmissions is Congestive Heart Failure (CHF) patients, which often have a readmit rate as high as 25%. In order to reduce readmissions, a hospital system must engage strategies which prevent readmissions. However there are strict limitations on the amount of expenditure which a hospital system may use to reach their goals. In order to prevent readmissions effectively and efficiently, we treat this as an optimization problem, in which we are required to complete all of the components outlined below.

- Prediction of readmission risk for Congestive Heart Failure patients
- Methods of interventions
  - Expected reduction in readmissions
  - Cost of Interventions
- Financial impact of Readmissions
- Optimization techniques to select the appropriate methods, and individual patients
1.2. Research Objectives

The objective of this research is to apply an analytics driven approach which combines several commonly used algorithms to each part of the readmission reduction problem. While the case considered is for CHF, it is trivial to apply this to other diagnosis groups. By giving the case of CHF, we are able to show that the framework developed is applicable in a real-world situation. We use a local community hospital system's records to provide the proof of concept.

1.3. Dissertation Overview

The dissertation is divided into six chapters, beginning with the initial introduction, and providing background information on the subject area. As this research lies in the intersection of many topic areas, we strive to address each, including the current state and trends of healthcare, and the description of the problem. Chapter 2 will include a summary of the problem as well as provide background information which is required for problem understanding it will also include a justification for this investigation and a literature review of other approaches which attempt to meet the same end goals. Chapter 3 will include an introduction into the data mining process which was used in the formulation of the risk model. As has already been discussed the creation of the risk model remains a central component to overall efficacy of this readmission reduction framework. We will discuss topics, which include the available data sources, the formats which data were received in data preprocessing steps, including the creation of new variables and transformations of previously existing variables. As well, information on the current modeling approach which is been shown to be successful will also be explained including data sampling methodologies prediction model, formulations and
classification results. With the prediction model explained Chapter 4 will focus on the information which would be available to other hospital systems or healthcare organizations for the development of their intervention programs. Within this proposal, the scope will be limited to defining the information available and explaining how this information could be utilized by healthcare systems. However, the end-use of this, as well as the formal definitions and methodology which should be used in transferring published literature into defined intervention strategies and cost assessments as well as assessments of overall efficacy will be left for future research. Chapter 5 will include information related to search heuristics in the overall development of the final intervention strategy for the company to utilize. We have already discussed how risk is assessed and predicted and also how literature provides insight into possible intervention strategies. This chapter will focus on how these two ideas can be condensed down and utilized in an analytic fashion to define an optimum strategy. The focus will also include those aspects of the organization structure, which would be important in the formulation of constraints as well as in the functions which might be considered or excluded. The remaining chapters will be focused on results, conclusions, and discussion related to the implementation of this readmission reduction framework.
2.1. Rise of Chronic Disease and the Readmission Crisis

At present, it is estimated that 75% of healthcare expenditures are focused on the treatment of chronic diseases.[1][2] A chronic disease, is one for which there is no known cure, which patients will have for the rest of their lives. While these conditions may be managed, reducing or eliminating the side-effects of the disease, there is no cure.

There are several factors at work which have lead to this. Improvements in healthcare can now prevent mortality in situations which would have previously been terminal. Improved chronic disease management programs lower the mortality risk of patients, and many other additional factors. However the explanation for this rise is of little interest to this investigation. We are merely concerned that this is the case, and what our aims should be since it is.

Having a growing proportion of the patient population living with one or more chronic disease has large effects on the healthcare system. One of the largest impacts deals with readmission rates. When a patient enters the hospital as an inpatient they are considered ‘admitted’ to the hospital. Once the patients’ treatment has completed, and they are stable, the patient is discharged. At this point, the goal is that the patient is no longer requiring hospital inpatient services, and should not require them for a long period of time. One metric for the quality of care that a hospital is providing is called the ‘readmit
rate’. Essentially a readmission occurs when the patient is once again admitted to the hospital after they have been discharged. Typically, the length of time which is of interest is 30 days. This means that a patient who is admitted to the hospital within 30 days of their last discharge is considered a readmission. However we should note that there are a variety of lengths which are often tracked, such as 7, 15, 60, 90, 365. The logic behind using this rate as a metric for hospital care delivery quality is that hospital systems which often have patients returning to the hospital after a discharge must be delivering inferior care as compared to those which are able to prevent a larger percentage of readmissions.

This also enables the elimination of a conflict of interest, whereby a hospital system would have an incentive to deliver poor care as that would generate additional patient volume from repeated admissions from patients.

The importance of this issue has been magnified recently due to an addition of section 1886(q) to the Social Security Act, an addition as part of the Affordable Care Act.[3] This act establishes several things. One is that hospitals with disproportionately high readmissions will face reductions in reimbursements. It does this by viewing the readmissions for patients with the conditions of Acute Myocardial Infarction (AMI), Heart Failure (HF) and Pneumonia (PN) and determining the excess readmission ratio, a calculation based on a national average and adjusting per the hospital's patient population, including aspects such as comorbidities, frailty, and demographic characteristics. In future, the conditions being targeted are likely to be expanded to include acute exacerbation of chronic obstructive pulmonary disease (COPD); and
patients admitted for elective total hip arthroplasty (THA) and total knee arthroplasty (TKA). The specific definitions for calculations are shown below.[4]

Formulas to Calculate the Readmission Adjustment Factor

Excess readmission ratio  =  \frac{\text{risk adjusted predicted readmissions}}{\text{risk adjusted expected readmissions}}

Aggregate payments for excess readmissions

= \left[ (\text{excess readmit ratio for AMI} - 1) \sum \text{base operating DRG payments for AMI} \right]
+ \left[ (\text{excess readmit ratio for HF} - 1) \sum \text{base operating DRG payments for HF} \right]
+ \left[ (\text{excess readmit ratio for PN} - 1) \sum \text{base operating DRG payments for PN} \right]

Aggregate payments for all discharges

= \sum \text{base operating DRG payments for all discharges}
Aggregate payments for all discharges

\[ = \left[ 1 - \left( \frac{\text{Aggregate payments for excess readmissions}}{\text{Aggregate payments for all discharges}} \right) \right] \]

Readmission Adjustment Factor

\[ = \begin{cases} 
\text{FY 2013: the higher of the Ratio or 0.99 (1\% reduction)} \\
\text{FY 2014: the higher of the Ratio or 0.98 (2\% reduction)} 
\end{cases} \]

Formulas to Compute the Readmission Payment Adjustment Amount

Wage adjusted DRG operating amount  =  DRG weight \( \times \) \((\text{labor share} \times \text{wage index}) + (\text{nonlabor share} \times \text{cola})\)

Base Operating DRG Payment Amount  =  Wage

\[ - \text{adjusted DRG operating amount} + \text{new technology payment} \]

Readmissions Payment Adjustment Amount  =  \([\text{Base operating DRG payment amount} \times \text{readmissions adjustment factor}] - \text{base operating DRG payment amount}\)
2.2. Problem Justification

As the above changes show, if a hospital system cannot reduce their readmission rate to a level which can avoid penalties, it will face a financial burden. However, whatever the strategy to reduce the readmission rate, a reasonable cost must be maintained, while still remaining effective. Otherwise penalties will still be applied, along with the cost of readmission reduction strategies. We see that this can pose a difficult problem to solve. Exacerbating the problem is a secondary issue which is prompted by the necessity of managing costs associated with readmission reduction programs. As the readmissions account for only a small percentage of the patients, the vast majority of patients require no interventions. As such, an initiative which targets all patients being discharged would be particularly wasteful. However that still leaves the issue of being able to identify those patients which would be likely to be readmitted. This is a difficult aspect of the problem as there are a multitude for reasons or characteristics which may be responsible for hospital readmissions among the patients, and quite likely differing ones for each of the separate diagnoses.

2.2.1. Optimization Problem

So we have a situation in which determining the optimum solution represents a non-trivial problem. We are faced with determining all of the following:

- What interventions to use
- How large should our intervention recipient group be
- Which specific individuals should be targeted
- What is the financial outcome
All of which are subject to several constraints in determining the solution for the optimum readmission reduction strategy. Consider the following constraints:

- Effectiveness of interventions (and changes based on reductions)
- Cost of interventions
- Additional staff required
- Staff duty reallocation
- Reimbursement reduction
- Improvements made by cohort hospitals
- Budget constraints

All of this must happen in an environment in which the use of experimentation to determine the strategy is largely imprudent. The outcomes of experimental trials would be on actual patients. Beyond this ethical aspect are the more practical problems associated with running experiments in which the variation between individual subjects is high, and the length of time required to generate sample sizes large enough to show significant results. Moreover, the managerial concerns of constructing an intervention program, even on a trial basis would be large. Also consider that a choice like this might very well require additional staff to accomplish. It would be problematic to go through a hiring process to implement a trial program, as the times to fill positions, especially skilled ones, can be quite long. Put simply, the readmission reduction strategy cannot be of the guess and check variety, but must be well thought out, analytics driven strategy, as
the timeliness of implementation is significant, and financial impact of these decisions is high.

What we propose in the solution of the problem represents a method whereby hospitals may leverage the existing literature from implemented readmission reduction initiatives, information which cannot be achieved internally, and combine that with internally derived analytics. From the information on effectiveness, we tie in the estimated financial costs associated with the specific hospital, and the readmission reduction which might be expected if a given strategy is employed. We utilize machine learning on the hospital patient information to formulate a risk prediction model which surpasses the accuracy of the statistically based models found in literature. At this point, the elements of the problem are assembled, and the remaining task is to define the optimization problem for the solution of which patients should be targeted to receive each intervention. Thus a hospital can generate an optimum strategy for the reduction of readmissions, all the while minimizing the financial losses due to reductions in reimbursements.

2.3. Literature Review

The literature review will cover the separate aspects of published literature which is connected to this problem. It examines two approaches which literature has used.

2.3.1. Reducing Readmissions

The current literature which focuses on readmissions is largely based in two separate directions. One centers around the intervention strategies. There are countless methods which have been trialed and analyzed. For CHF, these methods fall into multiple strategies. Some methods tried include multi-disciplinary teams[5], patient education based interventions[6], home monitoring[7], support[8] and transitional care[9]. There
are multiple types of interventions which have been attempted, as well as duplicates of many of the types. However what these papers describe is the intervention which was tried, the population which it was trialed on, and the results of the population. In rare cases, the scope is expended to incorporate the financial aspects such as the cost of interventions; however this is not always the case. For an intervention strategy to be effective, there must be an accounting for cost, for a hospital system which is attempting to utilized or implement these interventions or to develop an intervention reduction strategy must be assured that there will be some return on what is a costly expenditure.

In the other category are articles which examine the prediction of the readmission risk of patients. One study examined the efficacy of utilizing a bedside test for B-type peptide levels, as B-type peptide is secreted from the ventricles or lower chambers of the heart in response to pressure changes of patients [10]. Similarly, C-reactive protein was also examined [11]. Data mining techniques have also been aimed at the problem, in some cases aimed at creating prediction models. [12,13] Many examine which factors may be predictive of heart failure readmission, but do not propose models for risk assessment [14-15]. Some approaches have favored a multidisciplinary strategy in the effort to reduce readmission risk for CHF patients [5][16].

2.3.2. Predicting Readmission Risk

Work has been conducted on readmission risk prediction which utilized a logistic regression model.[17] However the results of the model are modest at best, yielding a C-statistic of 0.62. A typical point in which the model is considered useful if exceeding 0.7.[18] A value of 0.5 is considered random.[19] There are also multiple models which predict the mortality risk for patients with CHF [20][21][22][23][24], however there
exists an independence of correlates between the mortality and readmission [25].

In this problem domain, it is common to have a fairly small population, perhaps only a few hundred patients. Given the nature of hospital data collection, there will be a great deal of information collected on each of these patients. This makes it increasingly likely that there will be a high number of features in comparison to sample size. This is often referred to as high-dimensional statistics and occurs when the number of parameters $p$ is much larger than the number of samples $n$ [26]. Often, when problems become high-dimensional, there are a wide range of applicable techniques at the which span the gap between statistics and machine learning [27]. Regularized likelihood approaches also have been greatly utilized in this problem domain [26]. Linear models may still be parameterized for the equations of the form below,

$$Y_i = \mu + \sum_{j=1}^{p} \beta_j X_i^j + \epsilon_i (i = 1, ... n)$$

as long as the intercept $\mu$ and parameter vector $\beta$, if $p >> n$ as long as $\beta$ is sparse. And high-dimensional statistical inference leads to reasonable accuracy or asymptotic consistency if

$$\log(p) * (\text{sparsity}(\beta)) \ll n$$ [26].

One approach to dealing with the complexity present in new data sets was the use of boosting algorithms [28]. As well, novel methods were also created which could estimate the intrinsic dimensionality within data [29]. One common example of this is the LASSO which is used for variable selection in high dimensional data [30]. Improved versions of this idea would be SCAD [31] the Dantzig selector[32]. In the end these are variable selectors.
In application areas where very high dimensional data is the norm, such as in many aspects of Bioinformatics [33] or image processing [34], or digit recognition [35], the application of machine learning has become much more common [36]. However the application of these techniques to heart failure readmission is not well understood. However what these papers lack is what hospital systems truly need, to bridge the gap between having a knowledge of which intervention are possible, a description of the interventions strategies, and a method of prediction. For only when we are able to connect the three, is the problem bounded. If we are able to know how much an intervention costs, how effective it is, and who can receive it, then we are able to determine an optimum strategy which gives the proper intervention mix to the population.
CHAPTER III
RISK PREDICTION

In this chapter we provide the methodology followed in the creation of the risk prediction model. We begin with details about the data, which was available from the hospital system and describe the form and source of each variable. With this defined, we then describe the data preprocessing which was used to formulate the final list of possible variables available to us in the prediction of heart failure readmission risk. After the description of data as well as the preprocessing, we then follow with the modeling, both in creation and evaluation. As with most data mining projects or investigations, our research was largely iterative during the data modeling portion. We repeatedly cycled through data transformations as well as model tests and evaluations. We will describe both the methodology which was used as well as the general metrics used in the module evaluation. We describe the feature selection which was utilized in the formulation of the preliminary heart failure prediction model. The chapter concludes with a description of the final data preprocessing, outcomes of the feature selection and final model creation. Details are provided on the final model creation algorithm which was used and its source.

3.1. Data Collection

The data source is a local community hospital system. Data is from CHF patients who were discharged from that institution between May 2009, and August 2012. In the data set, there are a total of 652 patients. One of the important sources of information is the
case management assessment. This contains much of the information regarding social situation. This social information is one of the components which is often deemed relevant in the prediction of the patients risk, and this remains the only source of that information. However this information is not collected on every patient, only on a certain subset, based on the capability of case management to find time to complete it. This leaves complete information available on 452 of the 652 patients. The information which is available to us includes demographic information, billing and coding information, procedural coding information, encounter related information, as well as social and service based information.

Table 1: Overview of parameters used in data mining

<table>
<thead>
<tr>
<th>Procedures</th>
<th>Current Services &amp; Needs</th>
<th>Demographic</th>
<th>Hospital Stay</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thoracentesis</td>
<td>Living Arrangement: home</td>
<td>Patient Age</td>
<td>Hospital</td>
<td>Readmission within 30 days</td>
</tr>
<tr>
<td>Hemodialysis</td>
<td>Living Arrangement: Assisted Living</td>
<td>Patient Gender</td>
<td>Discharges Past 6 Months</td>
<td></td>
</tr>
<tr>
<td>Dx Ultrasound-Heart</td>
<td>Living Arrangement: Skilled Nursing Facility</td>
<td>MS-DRG</td>
<td>Discharges Past Year</td>
<td></td>
</tr>
<tr>
<td>Infusion Of Vasopressormonoyes</td>
<td>Living Arrangement: LTAC</td>
<td>APR-DRG</td>
<td>Discharges past Month</td>
<td></td>
</tr>
<tr>
<td>Insert Endotracheal Tube</td>
<td>Living Arrangement: Acute Rehab</td>
<td>Life With</td>
<td>Admit Type</td>
<td></td>
</tr>
<tr>
<td>Ven Cath Renal Dialysis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cont Inv Mec Ven &lt;96 Hrs</td>
<td>Hours w/o Assistance</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Venous Cath Nec</td>
<td>ADLsPTA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sm Bowel Endoscopy Nec</td>
<td>Ambulatory Baseline</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Packed Cell Transfusion</td>
<td>Adaptive Equipt: Cane</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Within the dataset, we find that the proportion of patients which were readmitted within 30 days is 23.8 percent, representing 107 patients. Within the group we have a 213/239 split of males and females. In the set of 452 patients, 105 were discharged to Skilled Nursing Facilities (SNF's), 124 were discharged with Home Health, 181 were discharged home with routine discharges, and the remaining 37 to hospice, rehabilitation facilities, Long-Term Facilities, General Hospital, and expired. A fairly large percentage, 21.2% had pneumonia, and 34.3% had hypertension.
While this information is readily attainable, literature predicates an interest in certain lab values which would be of interest prediction of a heart failure patient readmission risk. The most important would be the elevated b-type natrieretic peptide levels, (ProBNP).[10][37][38][39][40] Also of interest would be the hemoglobin A1c [41][42][43], as well as the creatinine levels.[15][44][45] However this information is not always available. Our interest in the patients would be for the most recent value, however if we limit ourselves to those cases which have the value recorded from that hospitalization, the proportion of missing values would be high. For this reason, we extend our period of inclusion to within 3 months of that discharge, taking the nearest chronologically to the discharge date. With this criteria established, we find that the patients with missing values represent 51, 313 and 3 patients for ProBNP, Hemoglobin A1c, and Creatinine respectively. In dealing with the missing values, we opted for mean substitution.

3.2. Data Cleaning

Recall that we found that the patients with missing values for ProBNP, Hemoglobin A1c, and Creatinine respectively. In dealing with the missing values, we opted for mean substitution. Although we note that Hemoglobin A1c will likely be unusable, as the majority of points are missing. Beyond this, the cleaning conducted on the dataset was primarily aimed at information from the care coordination assessment form. The cleaning required on the care coordination admission assessment requires additional work. Many of the fields allow for the entry of free text, requiring the categorization of the field into more similar entries. Largely this process is straight-forward, changing multiple methods of entering the same information into identical formats. For instance, a
common option for the field describing the patients’ living situation includes alone, family, or spouse. Consider for instance, one entry is ‘son, Gary’ which is altered to ‘family’. For those fields which have a simple change, for instance those fields which are missing or indicate not applicable, are filled with the most common value.

3.3. Data Mining

In the mining context, we will utilize several strategies. Initial aims focused on the classification rate of the classifier on the data set. The problem represents a classic classification problem, with a binary outcome. The prediction of the patients readmission is the end goal. Beyond this basic classification however, it is useful to determine the relative risk of each individual patient rather than a simple binary classification. This would allow for prioritization in the selection of patients for interventions, which is our main intent. The data we work with is largely nominal; however there are several continuous variables, as well as some which could be considered ordinal. For instance the lab values are certainly continuous; however one might consider the number of readmissions in the past six months to be ordinal, as there are only a few possible values with a natural order.

For the analysis of modeling techniques to apply we are faced with the question of prediction model performance. We may also extend the traditional method of comparison in data mining models. The schematic method which is presented in Giudici yields: criteria based on statistical tests, scoring functions, Bayesian criteria, computational criteria, and business criteria[46]. The method selected should fulfill the ultimate aims of the project, reducing the number of readmissions in the target conditions. In order to meet that aim, we are concerned with the classification
performance within subsections of the population, presumably beginning with the high risk populations. However we are not aware, and cannot determine the specific group sizes to which interventions will be applied. As such the target group is unbounded. Moreover we are more concerned with maximizing the true positive rate, the proportion of correctly labeled patients which have readmissions. This is sometimes referred to as sensitivity. However this rate is dynamic throughout the solutions space. When on the selected probability cut-off utilized is altered, so does classification performance. While in many instances, the point which maximizes the overall classification rate is used, this will not necessarily be the point used as the cut-point for interventions. In this way, we seek an effective risk profile, one which performs well throughout the solutions space, not at a particular point. Moreover we seek one which allows for preference toward smaller selections, or those solutions with stringent cut-off criteria, as many interventions are costly and not likely to be applied to a large percentage of the patient population.

With the preference toward small groups, a measure such as Area Under Curve (AUC) is not optimal. Our selection method for comparing data mining algorithm performance is the Receiver-Operator Characteristic (ROC) curve. This is a visual representation of the model performance through the solutions space. It depicts the view of performance under differing probability cut-off criteria.
The view of the ROC shown in the figure depicts a curve of performance. Largely the focus is on the determination of the correct modeling approach. We will be testing a large variety of data mining classification algorithms. The types of modeling approaches used are listed in the table below.

Table 2: Data Mining Algorithms trialed on the CHF readmissions data set

<table>
<thead>
<tr>
<th>Data Mining Algorithms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Support Vector Machines</td>
</tr>
</tbody>
</table>
3.3.1. Initial Investigation

In this problem domain we have very little evidence that one modeling or machine learning technique would give superior performance to any other. As has been shown previously in the literature review there has been little work on the prediction of heart failure readmission. In those scenarios which have presented both the modeling technique used and the results, all results have been of moderate success at best. Therefore to have certainty that the approach utilized in any machine learning practice for the prediction of heart failure readmission would give the optimal performance, we must attempt a multitude of machine learning algorithms. We have selected Rapid Miner 5.3 for the initial investigations as it gives the user the ability to test multiple machine learning algorithms quickly, as well as provides the ability to compare results between methods easily.
Rapid Miner 5.3 utilizes a flow-based interface as shown above. In order to attempt the comparison, we wish to view not only the ROC curve, but also the AUC performance. As well, we have selected a k-fold cross-validation. In k-fold cross-validation the data set is divided into k subsets. On each separate partition of the data a \( k^{th} \) part is held back for validation. The remainder of the data is used in the model creation. By iterating through each of the k partitions, we are able to get an averaged understanding of model performance.

Figure 2: Example of the Rapid Miner Process creation

Figure 3: Representation of the k-fold cross-validation which is often performed
Through this validation process we are able to avoid or at least constrain one problem of data mining, which is over-fitting. The flow, which was utilized to prepare and compare models, is shown below.

Figure 4: Rapid Miner process for predicting CHF readmissions through comparing ROC curves

The parameter selections allow further tuning of certain parameters. We have retained default settings in most cases. The process used follows the rudimentary data mining process whereby data is entered or drawn from the database, it then approaches some splitting criteria k-fold cross-validation. Sub-processes contained within the cross validation are shown below.

Figure 5: Sub-process of the compare ROC's operation, which gives the modeling algorithms and necessary transforms.
The results are combined and displayed through the compare ROC operator. As shown below, initial model performance is lacking in many respects.

Largely this is to be expected, although we can see that certain models do perform better than others. For instance, the random forest provides reasonably accurate results, although we should note that this is due to creation of a trivial solution. The solution which the tree defined utilizes the skew in the data specifically the minority and majority class differences in order to create a tree, which although achieving significant performance does little to add any insight from the data. The tree simply selects the majority class as the default guess in every scenario. In this instance the tree selects or predicts that each patient will not be readmitted as readmission is a rare case. In this way, the tree can we achieve reasonable performance although this is of little use.

Figure 6: Initial performance in readmission prediction
This particular problem is known as the class imbalance problem. A large amount of work in recent decades is gone to identifying certain methods or processes for dealing with the class imbalance problem. A class imbalance occurs often in medical diagnoses, predicting fraud and network intrusions. The problems faced in these situations are caused by the classifier being biased towards the majority class. Often in these situations, although the minor classes are rare, they hold greater importance.

There are three basic categories of research dealing with remedying this topic, the algorithmic approach, data preprocessing, and feature selection. Data preprocessing largely focuses on sampling techniques. Either new data points are added or removed from the overall data set when new entities are added. This is known as oversampling. When existing data points are removed this is known as undersampling. In the algorithmic approach the learning algorithm itself is modified. This often takes the form of cost sensitive methods or recognition-based approaches kernel-based learning. The support vector machine (SVM) and radial basis function (RBF) are examples of this approach. In feature selection a subset of the total number of features is used in the model creation. This is largely focused on high dimensional data sets. A filter is often used to score each feature independently based on some ranking criteria. In our attempts to explore our particular problem we utilize each of the approaches. Further details are provided below.

Figure 7: Entirety of a random forest tree for CHF readmission prediction simply guesses 'no' for every patient.
As the results show only reasonable performance, superior to a random guess, but lacking the performance which would be desired, the class imbalance problem is certainly present in the prediction of readmissions. This has been shown to hurt performance[47]. The strategies for remedying the results are three-fold[48]. In cases with large data sets, the majority class may be down-sized to the size of the minority class. However in some instances, such as ours, there is insufficient data to under-sample. In our case, the minority class is over-sampled. The final option ignores the classes by utilizing recognition based, rather than discrimination based inductive schema [49][50].

In our selection, over-sampling is the common route, however there are a multitude of methods available to accomplish this. The simplest is to simply select random points from the minority class with replacement until the classes are equal. However this has the drawback of homogenous groups, with many of the points adding little information. Slightly more complicated methods utilize the creation of points which did not originally exist in the original data set, but are generated using the information from it [51]. The option we select is Synthetic Minority Oversampling Technique (SMOTE) [52]. This algorithms creates points which are defined by using a nearest neighbors perspective, adopting features from those points nearby, but through the conglomeration of multiple points.

Figure 8: Representation of the SMOTE algorithm creating synthetic points
This oversampling method has consequences to its’ use. Before selecting this data preprocessing step for the modeling pipeline, we should first make sure that it is necessary. It has been shown that SMOTE performs significantly better after feature selection has been conducted, so we will present the results comparison in tandem with the feature selection [53].

3.3.2. Feature Selection

Many machine learning techniques were not designed to handle a large number of relevant features. As such, the necessity of feature selection techniques has been shown in many application areas. Feature selection techniques improve machine learning results in a number of ways. First and foremost, is that the removal of irrelevant features reduces overfitting and improves model performance. Another outcome of this reduction in features is faster and more cost-effective model building. As well it can also be useful in gaining deeper insight as to the role or nature of many of the variables in the model.

There is some certain trade-off in using feature selection techniques in the modeling process. This added complexity means another layer of decisions to make. Not only are we involved in machine learning for the prediction, but now we are also utilizing a new set of techniques in the removal of features. As such, the number of options which are available makes fully analyzing the problem more expensive. The search for the perfect model has another dimension in the hypothesis space. For classification, there are three main types of feature selection methods, namely filter, wrapper and embedded.[54] They each pose different advantages and disadvantages.
Table 3: Feature selection algorithms types

<table>
<thead>
<tr>
<th>Feature Selection Type</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Filter</td>
<td>Fast, scalable, independent of the classifier</td>
<td>Ignores interaction with the classifier</td>
</tr>
<tr>
<td>Wrapper</td>
<td>Simple, interacts with classifier, less computationally intensive than randomized methods</td>
<td>Higher risk of overfitting, more prone to local Optima, computationally intensive</td>
</tr>
<tr>
<td>Embedded</td>
<td>Interacts with classifier, includes feature dependencies, computational complexity better than wrapper methods</td>
<td>Completely classifier dependent</td>
</tr>
</tbody>
</table>

For this analysis, we have largely focused on filter methods, which prove more attractive for the abilities to be fast, independent of the classifier, and for scalability. The feature selection algorithms which have been selected for this study are shown in the table below. As well, the main detriment, the ignoring of interaction with classifier, is not as important as we are trying many types of classifiers in the study. As well, for the other methods it is not possible to try a multitude of combinations of feature selection technique and classification technique.

Table 4: Common filter feature selection methods

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variance Threshold</td>
<td>A filter technique which removes features with variance is under some threshold. At the extreme is zero variance, where all values are the same.</td>
</tr>
<tr>
<td>Univariate feature selection</td>
<td>Univariate feature selection works by selecting the best features based on univariate statistical tests. It can be seen as a preprocessing step to an estimator.</td>
</tr>
<tr>
<td>Tree-based Feature Selection</td>
<td>Using a tree learner which gives relative importance of variables based on the ensemble method</td>
</tr>
</tbody>
</table>
Recursive Feature Elimination

A model is run on the full data set, and the relative importance is at each feature is calculated. From that information, the worst performing feature or features is selected and eliminated from the data set. This process repeats itself till either all features are of reasonable importance or the maximum number of features to reduce has been reached.

For the preliminary results, we utilized a univariate feature selection, as well as a tree-based feature selection method. Univariate feature selection selects the factors based on univariate statistical tests. In our case for the classification problem, this is composed of chi-square tests as well as Kruskal-Wallis tests for the continuous variables. This reduces the variables to the top 32 variables.

### Table 5: Univariate Feature Selection for features often found in literature

<table>
<thead>
<tr>
<th>Feature</th>
<th>P-Value</th>
<th>Odds Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discharges Past 6 Months</td>
<td>0.794</td>
<td>1.02</td>
</tr>
<tr>
<td>Discharges Past Year</td>
<td>0.063</td>
<td>1.13</td>
</tr>
<tr>
<td>Discharges past Month</td>
<td>0.693</td>
<td>0.8</td>
</tr>
<tr>
<td>Previous AMIs</td>
<td>0.804</td>
<td>0.9</td>
</tr>
<tr>
<td>Creatinine Filled</td>
<td>0.015</td>
<td></td>
</tr>
<tr>
<td>ProBNP Filled</td>
<td>0.052</td>
<td></td>
</tr>
<tr>
<td>A1C Filled</td>
<td>0.402</td>
<td></td>
</tr>
</tbody>
</table>

### Table 6: Univariate statistical test feature selection for features found in the care coordination admission assessment

<table>
<thead>
<tr>
<th>Feature</th>
<th>P-Value</th>
<th>Odds Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lives With Coded</td>
<td>0.62</td>
<td>0.89</td>
</tr>
<tr>
<td>Number of Hours Alone Coded</td>
<td>0.033</td>
<td>0.71</td>
</tr>
<tr>
<td>ADL Coded</td>
<td>0.948</td>
<td>0.99</td>
</tr>
<tr>
<td>Ambulatory Baseline Coded</td>
<td>0.368</td>
<td>1.16</td>
</tr>
<tr>
<td>Adaptive Equipt: Cane</td>
<td>0.619</td>
<td>1.13</td>
</tr>
<tr>
<td>Adaptive Equipt: Walker</td>
<td>0.16</td>
<td>1.4</td>
</tr>
<tr>
<td>Adaptive Equipt: Wheelchair</td>
<td>0.273</td>
<td>0.64</td>
</tr>
<tr>
<td>Adaptive Equipt: BSC</td>
<td>0.439</td>
<td>0.52</td>
</tr>
<tr>
<td>Adaptive Equipt: Hospital Bed</td>
<td>0.682</td>
<td>0.62</td>
</tr>
<tr>
<td>Adaptive Equipt: Oxygen</td>
<td>0.005</td>
<td>2.26</td>
</tr>
<tr>
<td>Steps to Enter Home Coded</td>
<td>0.094</td>
<td>1.63</td>
</tr>
<tr>
<td>Bathroom Coded</td>
<td>0.409</td>
<td>0.53</td>
</tr>
</tbody>
</table>
Bedroom Coded 0.16 0.53
Current Services: HHC 0.128 1.52
Current Services: SNF 0.622 1.41
Current Services: Passport 0.999 0
Current Services: Meals on Wheels 0.693 1.42
Current Services: Pharmacy 0.174 0.72
Current Services: Lake County Council on Aging 0.874 0.83
Assess pt/family understanding d/c needs 0.181 0.66
No Questions 0.974 0.99
Additional needs identified 0.595 1.19

Table 7: Univariate statistical test feature selection for features obtained from Crimson care registry

<table>
<thead>
<tr>
<th>Feature Description</th>
<th>P-Value</th>
<th>Odds Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient Age</td>
<td>0.632</td>
<td></td>
</tr>
<tr>
<td>Gender Coded</td>
<td>0.116</td>
<td>0.71</td>
</tr>
<tr>
<td>MS-DRG</td>
<td>0.026</td>
<td>0.7</td>
</tr>
<tr>
<td>Severity Level</td>
<td>0.001</td>
<td>1.73</td>
</tr>
<tr>
<td>Mortality Risk</td>
<td>0.02</td>
<td>1.54</td>
</tr>
<tr>
<td>Hospital Coded</td>
<td>0.295</td>
<td>0.79</td>
</tr>
<tr>
<td>Admit Type</td>
<td>0.968</td>
<td>0.98</td>
</tr>
<tr>
<td>Case LOS</td>
<td>0.1</td>
<td>1.06</td>
</tr>
</tbody>
</table>

Table 8: Univariate statistical test feature selection from procedural and diagnosis coding

<table>
<thead>
<tr>
<th>Feature Description</th>
<th>P-Value</th>
<th>Odds Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hy Kid Nos W Cr Kid I-lv(403.90)</td>
<td>0.445</td>
<td>1.19</td>
</tr>
<tr>
<td>Thoracentesis(34.91)</td>
<td>0.946</td>
<td>0.97</td>
</tr>
<tr>
<td>Hemodialysis(39.95)</td>
<td>0.041</td>
<td>2.71</td>
</tr>
<tr>
<td>Dx Ultrasound-Heart(88.72)</td>
<td>0.562</td>
<td>0.53</td>
</tr>
<tr>
<td>Infusion Of Vasopressor(00.17)</td>
<td>0.579</td>
<td>1.62</td>
</tr>
<tr>
<td>Insert Endotracheal Tube(96.04)</td>
<td>0.579</td>
<td>1.62</td>
</tr>
<tr>
<td>Ven Cath Renal Dialysis(38.95)</td>
<td>0.579</td>
<td>1.62</td>
</tr>
<tr>
<td>Cont Inv Mec Ven &lt;96 Hrs(96.71)</td>
<td>0.399</td>
<td>2.17</td>
</tr>
<tr>
<td>Venous Cath Nec(38.93)</td>
<td>0.999</td>
<td>0</td>
</tr>
<tr>
<td>Sm Bowel Endoscopy Nec(45.13)</td>
<td>0.759</td>
<td>1.3</td>
</tr>
<tr>
<td>Packed Cell Transfusion(99.04)</td>
<td>0.149</td>
<td>3.29</td>
</tr>
</tbody>
</table>
The second analysis for feature selection was the tree-based feature selection method. It uses tree-based estimators which can create component feature importance is, and discards irrelevant features. When comparing the results achieved from this method of feature selection, to that of univariate feature selection we are interested in the impact that occurs on the predictive model performance. We are also interested in knowing which features are maintained in either method. When comparing model performance, all other aspects of predictive model creation have been held constant. The only thing that varies are model features.

Table 9: A Comparison of Features Selected in Each Method

<table>
<thead>
<tr>
<th>Feature Included</th>
<th>Univariate</th>
<th>Tree-Based</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diabetes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Prior History of HF</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Hemodialysis(39.95)</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Feature</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>------------------------------------------------</td>
<td>-----</td>
<td>----</td>
</tr>
<tr>
<td>Cont Inv Mec Ven &lt;96 Hrs (96.71)</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Packed Cell Transfusion (99.04)</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Chf Os (428.0)</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Atrial Fibrillation (427.31)</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Hypertension Os (401.9)</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Pneumonia</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Dmii Wo Cmp Nt St Uncntr (250.00)</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Creatinine Filled</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>ProBNP Filled</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>HourswoAssistance:Coded</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>AmbulatoryBaseline</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>AdaptiveEquipt:Walker</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>AdaptiveEquipt:Wheelchair</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>AdaptiveEquipt:Oxygen</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>NumberofStepstoEnterHome</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>BedroomLocation</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>CurrentServices:HHC</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>CurrentServices:Pharmacy</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Patientgender</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>MS-DRG</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Hospital</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Discharges Past Year</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Disposition</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>CaseLOS</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Univariate feature selection left a total of 27 features in the model. There were no features which were selected for inclusion by the tree-based model that were not in the univariate feature selection model. The tree-based model excluded an additional 12, leaving 15 features. A comparison of impacts of this type of feature selection on model performance is shown in figure 12. From the impact we can see overall that there is no difference to the mean ROC AUC. However, we do notice that the individual models
become less robust. There is a much higher variability in individual random forest of results than is seen in the results of the univariate feature selection. One explanation for this would be that a change was made to the number of features which are included when building the forest. This may mean that the feature selection under the tree-based provides a fewer number of features, but should still retain the same number of features in the building of each random forest.

3.3.3. Feature Selection Discussion with Domain Experts

One useful step in performing data mining is to confer with domain experts. Machine learning in general, relies on an entirely data-driven approach. Domain experts can provide an intuitive understanding of the reasons for the importance of each variable. Once the feature selection results had been compiled, the full listing was presented to hospital staff. The staff members were the Chief Quality Officer, a division manager for a medical floor, the director of Care Coordination, case managers, and social workers. On viewing the results, there was a general consensus that these results were expected. For instance they knew that patients who had many hours without assistance would be more likely to have their care neglected then those with additional assistance. In a similar way, patients who were discharged requiring home oxygen could often times return to the hospital very quickly if there were issues preventing the delivery to home. Having this discussion with domain experts allows us to take a certainty that there is an intuitive reasoning behind the variables which are used in the model. As well, one of our hospitals resides in a location which tends to see poorer patients, and it has been shown that economic factors certainly play a role in patient's health management. Aspects of the patient's previous history, especially their previous discharges plays a key role. If the
patient has required multiple hospitalizations, then they would be more likely to require them in future, and thus would be at an increased risk for readmission. So we can see that the information which we were able to determine as important via feature selection have an intuitive explanation as well.

3.3.4. Cost Sensitive Learning

In the algorithmic domain MetaCost was also used. Cost sensitive learning relies on imposing a penalty function or redefinition of weights to incorrect guesses[55][56]. In our example, this would take the form of providing a harsh penalty on the machine learning algorithms accuracy for incorrectly guessing that a readmitted patient would not be readmitted. It accomplishes this by wrapping any previously specified classifier in a cost minimizing procedure[57]. This is beneficial as it treats the underlying classifier as a black box and can be applied to any classification method. It also incorporates a variant of the traditional bagging ensemble method. In a bagging procedure, a training set is resampled through bootstrapping whereby new samples are constructed with replacement from the original sample[58][59]. This procedure is repeated and models are aggregated by uniform voting. MetaCost differs from traditional bagging by allowing the sample set which is drawn in the bagging procedure to be smaller than the original full set. MetaCost works by forming bootstrap replicates of the initial training set and creating a classifier with each set [60]. An estimate of each classes' probability in each example is calculated by the ensemble voting. Each training example is relabeled with the estimation for optimal class and the classifier is reapplied on the relabeled training set [61].
Overall, the results are largely unchanged by MetaCosting. In this example appears to be less useful than oversampling methods. The figure below displays the classification performance through ROC curves generated under MetaCosting.

Figure 9: ROC curve obtained through data post feature selection and oversampling with an optimistic bias
3.3.5. Selection of Modeling Approach

While the exploratory machine learning has been conducted in the Rapid Miner 5.3 environment, the software is proprietary, and the specifics on the machine learning algorithms used are not available to us in the search for a final package for utilization. One of the key aspects would be the ability to incorporate the final machine learning model into a distributable package for the healthcare system to utilize in daily workflow[62], [63]. In order to accomplish this, the language used should be flexible, extensible, and provide the ability to interface with multiple operating system distributions. As well, it would be beneficial if the specific packages and language used included sufficient documentation and examples.

Figure 10: ROC curve obtained through data post feature selection and oversampling with an pessimistic bias
The final selection for the modeling environment the Python programming language [64]. We also utilize Scikit-learn, a Python module includes a large variety of state-of-the-art machine learning algorithms [65]. Largely this module was intended for small-scale and medium scale problems focusing on both supervised and unsupervised problems. One of the greatest benefits which Scikit-learn provides is the excellent documentation, as well as API consistency. This Python module is distributed under the simplified BSD license and the source code and binaries are freely available. Scikit-learn offers an additional benefit in the form of the inclusion of compiled code. This can offer a significant speed advantage in many machine learning tasks. As well, it limits the dependencies required to only numpy and scipy [66]. The documentation is extensive with over 300 pages of user guide material and narrative documentation. Beyond this class references and tutorials are also provided.

In the final comparison, we utilize k-fold cross-validation and display ROC curves with AUC given as the metric of choice in comparison. The random forest is utilized as a learning algorithm of choice. The random forest utilizes 10 trees and the cross validation is conducted with the stratification and six folds. The data are oversampled using SMOTE.
Figure 11: ROC curve for the prediction of CHF patient readmissions under univariate feature selection.

Figure 12: ROC Curve for patient readmissions under the features selected by the tree-based feature selection.
Figure 13: Performance of the random forest on the classification of CHF readmissions with SMOTE

Figure 14: Performance of the random forest on the classification of CHF readmissions without SMOTE
3.3.6. Calculating Patient Risk

Our main interest in the creation of the predictive model is the calculation of patient risk. We wish to ascertain the probability of a patient being readmitted. The specifics of how this probability is determined will now be discussed. The model we are using in this case, it is a random forest classifier. This classifier is made up of an ensemble of decision trees. This model is capable of generating either the final classification for a patient, or also the class probabilities. The predicted class probabilities are determined based on the mean of the predicted class probabilities of the decision trees within the forest. Each of the trees calculates the class probability based on the probability distribution of samples in the trees’ and leaf node. To show this more clearly, we’ve constructed a small example.

Consider a random forest which had only three decision trees. The decision trees are used to predict the probability of patient readmissions. In the following figure, we see the final leaf nodes on each tree that a given patient has fallen into.

![Final leaf nodes of an example patient](image)

**Figure 15: Final leaf nodes of an example patient**

Each tree gets an equal vote. We take the average of the three trees that are in our example random forest. As a result, the probability of our readmission for this patient would be estimated at the average of 67%, 90% and 50%. This gives the final result of
69% probability that the patient will be readmitted. This risk score will come into play in the optimization algorithm discussed in Chapter 4.

3.3.7. Finalization of Modeling Approach

In the finalization of the modeling approach, we need to make final decisions on several different aspects of the modeling pipeline over which we have options. Each of these has been discussed in the previous sections. The aim in developing this predictive model is to show two things. The first is that it is possible to create a model with sufficient predictive power from hospital data. The second is to show the impacts of this predictive model with respect to developing a readmission reduction strategy.

3.3.7.1. The Class Imbalance Problem

In attempting to solve the class imbalance problem we tested a number of approaches. Attempts at MetaCost were unsuccessful in alleviating the problem. As well, the small number of patients in our data set preempted our use of under sampling. Oversampling utilizing SMOTE offered our best performance gains. It is understood that oversampling with this method will impose some overfitting during the learning process. However the trade-off to re-incentivize learning on the minority class seems to be more potent.

3.3.7.2. Learning Algorithm

Of the learning algorithms tried, the random forest provided the best results. It is unlikely that the same learning algorithm would yield the best results for all diagnosis groups, and for all hospital systems. It may be possible that additional investigation would yield a model which performs better than the random forest created, however we limit our investigation to this model as it has provided reasonable results.
3.3.7.3. Feature Selection

Of the two methods tried, tree-based feature selection provided a model yielding comparable results, but with a much lower number of variables. For this reason we have adopted the use of tree-based feature selection.
CHAPTER IV

OBTAINING INTERVENTION OPTIONS FROM LITERATURE

In this chapter we will discuss the current work which has been conducted on using literature on intervention strategies to formulate possible interventions which should be considered. We will discuss the aspects of the literature which are necessary for inclusion. Beyond this we will also investigate how the literature can be condensed down into simple aspects in order to provide a like to like comparison between published works. Details on the requirements of each facet are provided, as well as methodologies which may be employed in order to define the constraints which will be used in further work.

As already discussed, one of the largest hurdles a healthcare organization will face when developing their intervention reduction strategy will be to define the possible options which may exist. There is a wealth of information available on intervention strategies as well as many which define the effectiveness of the overall strategy. As well, it is not difficult to take the descriptions available from literature for different interventions and their effectiveness and to translate those into a system cost or intervention program cost, either as a whole or on the per patient level. However, what does remain difficult to do is to formulate which possible interventions could be used. Also work must be done to consolidate and give a level and fair comparison between the published works on different intervention strategies.
As has already been discussed, it is problematic to treat intervention strategies as an experiment. The amount of time which would be taken in designing how that strategy would be delivered, combined with the cost of the interventions themselves would be troublesome. Moreover, the actual implementation of an intervention reduction strategy likely requires an investment in capital largely human capital. It is unlikely that any healthcare organization will have the required number of staff members, especially those of the correct training and licensure available to assist on any implementation strategy on even a trial basis. As well, consider that the very nature of the readmission reduction problem is human centered. And as it has human subjects, it is difficult to get a reliable response. As such, it will simply require a larger sample size and a larger investment from the organization. It is simply not feasible for any system to attempt to address this problem in experimental manner.

Table 10: Relevant features which must be answerable by an intervention article

<table>
<thead>
<tr>
<th>Intervention Component</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>What</td>
<td>What is done for patients in this intervention</td>
</tr>
<tr>
<td>Who</td>
<td>Who is selected for inclusion</td>
</tr>
<tr>
<td>How Many</td>
<td>How many patients are selected</td>
</tr>
<tr>
<td>Who</td>
<td>Which staff members are conducting which parts of the intervention</td>
</tr>
<tr>
<td>Outcome</td>
<td>What were the results of the implemented trial</td>
</tr>
<tr>
<td>How often</td>
<td>What schedule was used for delivering interventions</td>
</tr>
<tr>
<td>When</td>
<td>When are interventions stopped</td>
</tr>
</tbody>
</table>
This largely leaves one single approach in determining interventions. Published literature was created through trials in differing healthcare organizations of different intervention strategies. The information published details the population to which the intervention was applied, as well as the description of the intervention itself. It will also likely contain the outcomes of that intervention strategy. By combining the insight and information from multitude of studies and trials conducted by different organizations one can gain a deeper and wider insight into possible strategies. There are some aspects of this formulation which are nontrivial. There are several components which are necessary for an organization to define. For each paper to be considered, it must be able to define all of these aspects. These are outlined in the table above.

The first and foremost would be a definition of the population to which the intervention strategy was applied. For there to be some measure of understanding of the efficacy of each intervention strategy there must be a comparison group, a control group. Largely, this presents no issue as each paper encountered thus far in research has provided this information. However, we do know that without this information, we would be unable to assess or assign the expected readmission reduction which the strategy is responsible for.

The description of each intervention strategy remains a central component. For the understanding of and implementation of the strategies defined must provide enough information to assess which patients were included, which patients were treated and when. As well, the description should provide details on how the patient is treated. It is also beneficial to define which staff members were responsible for which duties or, failing this, to at least describe the duties which are performed such that an organization
could define which staff members would have the requisite training or licensure to accomplish the task. It would also be beneficial to understand the length of time which is required for each intervention. This could be defined either from the patient perspective, giving details about the amount of interaction the patient has with the healthcare system. It would also be possible for the literature to define the amount of caseload or daily workload each team member was responsible for overseeing. For instance, they may describe that a case manager has a panel size of 63 patients. It should describe when the intervention is begun. Often this is immediately upon discharge, or perhaps before discharge has happened. It should also describe how the intervention strategy ends.

4.1. Efficacy

The primary focus, which must be considered, is the end goal of efficacy. How effective would this solution be if applied in our organization? While cost is of paramount importance, we cannot forget that the end goal is reduced readmissions or the effectiveness of our readmission intervention strategy. However, we are faced with some challenge as we try and navigate between published literature and an understanding of the readmission reduction we would see within our organization. One problem which largely cannot be overcome is an understanding that interventions may function differently on a personal level. We can understand that each individual patient may have a differing root cause of their readmission. For instance, the reasons which are responsible for their readmission may be related to noncompliance with diet. However, it is just as possible that they are related to noncompliance with medications, noncompliance with physician visits, or with daily self-management. With all of these differing aspects of human nature, how should we then address each intervention to this broad population? It would
seem that it would be best to have a detailed interview with each patient in an attempt to find or define the root cause [67]. But consider that the attempt to find the root cause would be itself an intervention [68–71]. Some specific types of readmission reduction strategies focus on finding the root cause, often through the use of a case manager, or social worker involvement with the patient. These individuals will conduct interviews and attempt to find or provide resources which are missing or needed for the patient. However it is an intervention which only provides information, perhaps actionable, but with no value contributed in itself. It requires further work beyond that already expended. It would seem that to reduce readmissions we must find and address the root cause for every patient.

Table 11: Outcomes of Literature, Utilizing Patient Interviews On Reducing Readmissions

<table>
<thead>
<tr>
<th>Title</th>
<th>Readmission Reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Telemonitoring in patients with heart failure</td>
<td>0%</td>
</tr>
<tr>
<td>Effect of a standardized nurse case-management telephone intervention on resource use in patients with chronic heart failure</td>
<td>46%</td>
</tr>
<tr>
<td>Case management in a heterogeneous congestive heart failure population: a randomized controlled trial</td>
<td>0%</td>
</tr>
<tr>
<td>Reasons for readmission in heart failure: perspectives of patients, caregivers, cardiologists, and heart failure nurses</td>
<td>Not reported</td>
</tr>
<tr>
<td>A case manager intervention to reduce readmissions</td>
<td>0%</td>
</tr>
</tbody>
</table>

The table above shows the outcomes from several of these programs. We can see that Largely, the effect of these programs was minimal. The one exception which achieved a readmission reduction of 46% utilized telephone interventions, which go beyond a limited case management interview for the determination of root cause.
However, literature does show that those investigations which do target a large panel size, or which do target a large portion of the overall CHF population with the rather broad tactics do see a significant reduction in readmissions[72]–[75]. In this we can say while it is certainly true that not every intervention will prevent every readmission. Nor is it possible to say that we must choose the correct intervention for each individual patient. We are at some point left with taking a larger viewpoint in the population as a whole. For this reason we are able to limit ourselves to the assessment of risk, rather than being required to predict the root cause. In this way, it is not necessary to focus on attempting to find the root cause, or in attempting to predict what the root cause would be.

This does leave an altogether different problem which must be considered. As each of the panel sizes or populations to which the interventions are applied are different, we must then define how to transfer each of the strategies effectiveness to a population which would be of our own choosing. For instance, consider an intervention which is applied only to patients of a high severity, or only to all population of advanced age. If we instead took that population and redefined the group as our entire population, or rather only the low risk portion of our population, we could very well expect differing results.

The easiest method for navigating this transformation would be to simply define the outcomes seen on that specific population with in the hospital system's own history. By using historical data, the hospital system would then be able to define an approximation of the effectiveness on their own group.
4.2. Methodology Followed

To demonstrate the proof of concept, we begin by defining the literature search we have covered in table 1. If each one of these aspects is not covered in enough detail for the specifics of the intervention strategy to be implemented in an organization with reasonable hope of similarity, then it is not possible to use that paper. We first do a literature search for readmission reduction interventions applied to the diagnosis group we are interested in, in this case, CHF. While the scope of literature in this incident is technically finite, it will largely provide more information than is possible to condense manually. As such, we are limiting ourselves to a small sampling of the relevant literature.

In the end, 62 papers were found for inclusion in the study. We follow the same process which would be used by an individual designing the possible readmission reduction programs. Each of the 62 articles is grouped into a certain type of readmission reduction program. Table 10 shows the general groupings for the intervention types. As well, it also gives the number of articles which were considered in each group.

Table 12: Amount of Literature by Group

<table>
<thead>
<tr>
<th>Readmission Program Group</th>
<th>Number Of Articles Considered</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case management interviews</td>
<td>5</td>
</tr>
<tr>
<td>Disease management programs</td>
<td>9</td>
</tr>
<tr>
<td>Education programs</td>
<td>7</td>
</tr>
<tr>
<td>Home care programs</td>
<td>9</td>
</tr>
<tr>
<td>Medication management programs</td>
<td>3</td>
</tr>
<tr>
<td>Multidisciplinary teams</td>
<td>8</td>
</tr>
</tbody>
</table>
Each article is first read for compliance to the specific criteria which are outlined in table 8, for inclusion in the study. The process is relatively straightforward, and examples are provided below for individual articles. It walks through the thought process, which is used to determine if a paper should be removed from consideration. A brief description of each article is provided, and then an attempt to create the table of necessary information is been conducted.

4.2.1. Example 1

The first article which we are going to use as an example for condensing literature into a concise framework for cross comparison purposes, covers a multi-disciplinary intervention which is largely nurse directed. We can see that it becomes very easy to answer the required information based on published literature. In the table below, we answer all the different components which are necessary for inclusion. From this information, the cost of providing this type of intervention strategy could be determined by an organization.

Table 13: a synopsis of the relevant features covered in an article

<table>
<thead>
<tr>
<th>Intervention Component</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>What</td>
<td>Nurses directed the interventions for comprehensive education to patient and family including dietary information, social service consultation and planning for discharge and review of meds. It also includes an intensive follow-up</td>
</tr>
</tbody>
</table>
Patients included were 70 years of age with a diagnosis of heart failure and had a risk factor of prior history of heart failure, four or more hospitalizations in five years, CHF, precipitated by AMI or uncontrolled hypertension, or diastolic blood pressure greater than 105mm Hg.

94 admissions in the control group and 53 in the treatment group

Nurses, registered dietitians, social services, geriatric cardiologist

59 patients in the control group had at least one readmission, and 41 patients in the treatment group had at least one readmission

7.2 hours of nursing time for patient, with each patient receiving one dietitian created meal plan, one social services, discussion, and one geriatric cardiologist medication reconciliation.

Interventions were stopped after 90 days.

4.2.2. Example 2

For our next example case, we look at a nurse case management, telephone intervention.[77] while this type of intervention is slightly different than the previous version, it is still just as simple to condense into the reusable framework.

Table 14: a synopsis of the relevant features covered in an article

<table>
<thead>
<tr>
<th>Intervention Component</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>What</td>
<td>In this study, the intervention group was telephoned within 5 days after hospital discharge and thereafter at a frequency guided by the software and case manager judgment based on patient symptoms,</td>
</tr>
</tbody>
</table>
knowledge, and needs. Patients received an average of 17 phone calls at decreasing levels of intensity, length, and frequency over the 6-month follow-up period.

**Who**
Patients were included if they had a confirmed clinical diagnosis of heart failure as the primary reason for hospitalization, and spoke either English or Spanish. Exclusions were made for patients with a psychiatric illness, severe renal failure, terminal disease discharge to long-term care facility, or previous enrollment in a heart failure disease management program.

**How Many**
130 patients were included in the intervention group, and 228 were included in the control group.

**Who**
Nurses trained in case management.

**Outcome**
31% readmission in the control group, and 17% readmission in the intervention group.

**How often**
The average patient required 16 hours of case management over a six-month period.

**When**
Interventions were stopped after six months.

### 4.2.3. Example 3

The third example that we cover is a home-based intervention, which unlike the previous two which is focused on hospital staff conducting these interventions at the institution itself, these interventions are conducted at the patient's own home.[78] The condensed version of the relevant features of the article is covered in the following table.
Table 15: a synopsis of the relevant features covered in an article

<table>
<thead>
<tr>
<th>Intervention Component</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>What</td>
<td>The home-based intervention includes a nurse and a pharmacist team, which optimizes medication management and identifies any clinical deterioration and intensifies medical follow-up and caregiver vigilance as appropriate.</td>
</tr>
<tr>
<td>Who</td>
<td>Patients were included if they had a hospitalization with congestive heart failure and impaired systolic function, along with an intolerance to exercise and a history of one or more hospital readmissions for acute heart failure.</td>
</tr>
<tr>
<td>How Many</td>
<td>There were 48 patients in the control group, and 49 patients in the treatment group.</td>
</tr>
<tr>
<td>Who</td>
<td>Nurses, pharmacists</td>
</tr>
<tr>
<td>Outcome</td>
<td>The patient receives one single home visit from the nurse and pharmacist team.</td>
</tr>
<tr>
<td>When</td>
<td>Interventions were stopped after six months</td>
</tr>
</tbody>
</table>

In this paper, there is no listing for the percentage of patients in the control and home-based intervention groups readmission percentages within 30 days of discharge. As this information is missing from the study, this article cannot be used as a possible intervention to consider. Although it describes the rest of the information which would be necessary to implement, we can have no understanding of its impact on 30 day readmissions.
4.3. Summarized Literature on CHF Readmission Interventions

The same process is applied to all literature which has been compiled. Overall, the largest change was with respect to the readmission program group ‘other’. For this group, all 19 articles have been removed. The reason for this is largely that these types of investigations were not limited to a commonly used type of readmission intervention, and far less likely to follow or provide the information typically found in these articles. Considered overall, and excluding the category of ‘Other’, we can see that approximately 60% of literature on readmission reduction interventions will not possess or provide all of the useful information. One of the main reasons for exclusion was that the program did not look at or provide information on their readmission reduction rate in the control group and in the intervention group. If this information is not provided, regardless of the detail provided on the interventions and outcomes with respect to mortality, self-care, diet adherence, or overall health care costs, we are not able to ascertain the specific impact on readmission rates. As we are missing this information, we cannot include this as a possibility.

Table 16: Amount of Literature after Validity Assessment

<table>
<thead>
<tr>
<th>Readmission Program Group</th>
<th>Initial number of articles</th>
<th>Final Number Of Articles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case management interviews</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>Disease management programs</td>
<td>9</td>
<td>5</td>
</tr>
<tr>
<td>Education programs</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>Home care programs</td>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>Medication management programs</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Multidisciplinary teams</td>
<td>8</td>
<td>2</td>
</tr>
</tbody>
</table>
And all of the remaining articles, we have the ability to define all the necessary aspects of the readmission reduction program with enough detail, but it could be replicated within an organization. Also, given that we have that level of detail, we can also establish an estimate for intervention program cost on the per patient side, and with respect to necessary startup investments.

4.4. Formulation of Estimates for Experimentation

The final number of articles represented in the search is still quite large. We also have multiple articles which offer of very similar approach. It does not make intuitive sense that we should have two separate programs which are very similar. When similar programs are compared, we take that which offers the superior results. This leaves us with eight separate program examples.

Table 17: Finalized List of Eight Programs Considered for Heuristics

<table>
<thead>
<tr>
<th>Program #</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Impact of a comprehensive heart failure management program on hospital readmission and functional status of patients with advanced heart failure [79]</td>
</tr>
<tr>
<td>2</td>
<td>Reduction in heart failure events by the addition of a clinical pharmacist to the heart failure management team: results of the Pharmacist in Heart Failure Assessment Recommendation and Monitoring (PHARM) Study [80]</td>
</tr>
<tr>
<td>3</td>
<td>Discharge education improves clinical outcomes in patients with chronic heart failure [74]</td>
</tr>
<tr>
<td>4</td>
<td>A multidisciplinary intervention to prevent the readmission of elderly patients with congestive heart failure [76]</td>
</tr>
<tr>
<td>5</td>
<td>Nurse-led heart failure clinics improve survival and self-care behaviour in patients with heart failure Results from a prospective, randomised trial [81]</td>
</tr>
</tbody>
</table>
The final eight articles selected come from a variety of the general intervention groups. From the information published in each of these articles, estimates are created for the amount of labor which would be spent on each patient. The estimates developed include the total hours spent in the 30 day period of time. The full listing of all roles and all estimates are provided in the table below.

Table 18: Finalized List of Eight Programs Considered for Heuristics

<table>
<thead>
<tr>
<th>Program</th>
<th>Effect</th>
<th>Nurse Hours</th>
<th>SW Hours</th>
<th>Pharm. Hours</th>
<th>M.D. Hours</th>
<th>C.M. Hours</th>
<th>Resources</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>85%</td>
<td>3</td>
<td>0</td>
<td>0.5</td>
<td>4</td>
<td>0</td>
<td>6000</td>
</tr>
<tr>
<td>2</td>
<td>73%</td>
<td>0.5</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>65%</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>56%</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>55%</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>47%</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>46%</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>13%</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

The hourly rates which are used are based on the average hourly wage paid for these positions within our organization. The last column, resources represents the initial startup capital which would be necessary to purchase systems the hospital system does not already own.
4.5. Discussion of Results

We have shown that it is relatively simple to take an initial literature review, which may include a very large volume of papers, and to condense them all down into the relevant information for selecting or prioritizing of CHF readmission reduction strategy. The same processor framework can be reused easily for any other diagnosis. The main aim of this type of framework is to give some constraints and boundaries for the selection of our readmission reduction strategy.

One primary benefit of this framework is not necessary for the intervention method literature to state the cost. Since we are defining the scope or the usage of this framework as beneficial to a single institution, the costs associated with specific intervention can be specifically calculated for that institution based on the work performed. This also allows us to not limit our cases to certain localities, or certain time periods. It is well known that the cost of a program would be largely determined on staffing costs, which vary by time and by location.[83][84] . This also provides an institution with more certainty about the specific costs, as these are now location dependent. It also allows the hospital system to take into account other factors that could very well play crucial role in startup costs, such as availability of current staff hours, or whether office space is available for new staff. These are specific facts which could only be answered by the institution, which is going to be performing that readmission reduction intervention.

The implementation of this framework is not limited to a random selection of patients. In fact, we can achieve even greater results by prioritizing which patients should be selected for interventions. In order to truly treat the readmission reduction problem analytically, we need three main components. We need the framework for condensing literature of
intervention options which have been attempted. This gives us a defined set of options for our organization. We also need some method of prioritizing patients. This can be done through predicting the risk of readmission for the diagnosis. Once we have these two things defined, we can then move on to the financial optimization problem. We may utilize heuristic searches in order to define the intervention or interventions, as well as the specific groups to which these interventions should be applied.

By treating the entirety of the problem in an analytical fashion, we can begin to face a downward pressure on readmissions for patients, but do it in a way which is fiscally sound. The way that we have constructed the framework makes it usable for any diagnosis type, and allows for a simple and straightforward method of determining which intervention strategies would be possible.
CHAPTER V
OPTIMIZATION OF IMPLEMENTATION THROUGH HUERISTICS

The genetic algorithm is an optimization heuristic which requires two parts to be implemented [85]. They draw from evolutionary computing, a subset of artificial intelligence. The first is a genetic representation of an individual solution to the problem. The other is a method for evaluating that solutions effectiveness.

The algorithm begins with the generation of several possible solutions. These represent the starting generation. From this point, the solutions will be evolved from this group iteratively. Each parent generation will create a child generation through reproduction, which is composed of two mechanisms. The initial reproduction is often referred to as crossover. A representation of this can be seen in the following figure. Beyond this combination of the features or genes of the parents is a subsequent mutation. By this mechanism, the solution can be created from the genes of the previous generation; however it will still possess the ability to shift in new directions in the solutions space which have not been explored previously.
Figure 16: Genetic algorithm which was implemented to determine the optimal solution for applied interventions.
The method for representing possible solutions genetically is very much similar to the formation of a vector of numbers. For this work in which we compare four intervention strategies, and using a sample of 300 patients. The solution space can be represented by a vector which is 8 units long. To define a solution, we are essentially looking at which patients are receiving which interventions. In order for this to work effectively, it is necessary to arrange the patients themselves into a vector which is increasing or decreasing in risk probability. Thus treatments will be assigned contiguously, thus targeting a population subset which will be between two risk scores. A representation of this idea can be seen in the figure below.

![Figure 17: Representation of the two point crossover mechanism which was utilized](image)

Figure 17: Representation of the two point crossover mechanism which was utilized

![Figure 18: Patient risk is laid out sequentially, with patients at highest risk first.](image)

Figure 18: Patient risk is laid out sequentially, with patients at highest risk first.
In order to define the solutions space, we need to have the definition for each of the intervention strategies. To define this, we are able to use two measurements, the start point, and the length of the vector. This is done for each intervention strategy resulting in 8 units for the vector, as seen below.

Individual: The Genetic representation is shown below

Description of each gene in individual
a: Starting point of intervention 1
b: Length of intervention 1
c: Starting point of intervention 2
d: Length of intervention 2
e: Starting point of intervention 3
f: Length of intervention 3
g: Starting point of intervention 4
h: Length of intervention 4

Figure 19: Representation of the solution vector format used

The problem which we are focused on defining the results based on $p$ patients, which will be receiving some or none of the $i$ interventions. Each patient will have a risk of readmission $\beta_p$. The risk reduction associated with the $i^{th}$ intervention is $\gamma_i$. The cost of a given intervention is $c_i$ and the resulting risk post intervention would be $\epsilon_p$.

When we consider the reduction in readmission risk, it would seem that the results of multiple interventions should not be additive. For example, a patient $p$ who has risk $\beta_p$ should after having intervention 1 and 2 applied would not expect

$$\epsilon_p = \beta_p - (\beta_p \cdot \gamma_1) - (\beta_p \cdot \gamma_2)$$

as it is not plausible that applying multiple interventions to a single patient will achieve the full reduction from both as there is often some overlap between interventions. This
being the case, it is not plausible that there would be no additional benefits to multiple interventions. As such, we utilized a method whereby there is an exponential reduction in the patients risk reduction benefits. For this study, we use a degradation factor $\delta_i$ such that

$$
\delta_1 = 1 ; \quad \delta_{n+1} = \delta_{n}/2
$$

Therefore we may define

$$
\epsilon_p = \beta_p - (\beta_p \cdot \delta_1 \cdot \gamma_1) - \left(\beta_p \cdot \delta_{1/2} \cdot \gamma_2\right) \quad \text{with} \quad \delta_1 = 1
$$

therefore

$$
\epsilon_p = \beta_p - (\beta_p \cdot \gamma_1) - \left(\beta_p \cdot \frac{1}{2} \cdot \gamma_2\right)
$$

with

$$
\gamma_1 > \gamma_2 > \gamma_3 > \cdots > \gamma_i
$$

which gives the reduction in the efficacy is applied to those interventions which are less effective than others. For instance if one intervention has a reduction of 50%, and another 25%, a patient sill receive the full reduction from the 50%, but will only receive a 12.5% reduction from the second.
A valid solution is one such that in solution vector $V$ of length $n$

$$
\begin{align*}
  V_n &< \psi; \text{ for odd } n \\
  V_{n-1} + V_n &< \psi; \text{ for even } n \\
  V_n &\geq 0 \text{ for all } n
\end{align*}
$$

Our goal is the overall goal is the calculation of overall cost, being that the optimization of overall financial return $\varphi$. In this way, we define that the overall financial impact is the aggregate case charges, assessment of the cost of interventions, cost of readmissions, but saving on readmissions prevented. In order to calculate this quantity, it is necessary to represent the case charges for each patient $\psi_p$ which is able to be transferred into the readmission adjusted payment amount $\chi_p$ by the transforms in equation 1 through 8. Note this adjustment causes a zero return for readmissions $\chi_p'$. With the conclusion of the methods for determining the prediction method, as well as the method which will determine the optimum distribution of interventions amongst patients a hospital or hospital system is able to expend resources to the prevention of readmissions while maintaining profitability. Without the understanding which is brought by having prediction, literature on interventions, and implementation heuristics, there is little possibility of success.
5.1. Understanding the Genetic Algorithm

We begin with the full population of all CHF patients. Each of these patients has a risk which has been assigned by the learning algorithm in Chapter 3. As well, we also know the outcome for that patient. It is important to include the actual outcome of the patient in the optimization heuristic. If we simply used the risk of the patient which is predicted, then we have no method of degrading optimization performance based on our predictive model performance.

All of the patients are arranged based on the risk score of the patient. If patients have the exact same risk score, they are randomly placed in order. A sample of 300 patients is taken from the initial population with replacement. We now have a vector with 300 patients, ordered by risks, and with known outcomes.

The goal of the program is to evaluate the impact of specific program interventions on this example population. We need to calculate the risk reduction faced by each patient. This can be done by parsing the chromosome which is used in the genetic algorithm. Recall that the chromosome includes all start points, and lengths for each of the program options. In order to do this, an array which details the interventions must also be created. This array contains the necessary information about each intervention. It includes the effectiveness, the amount of hours which would be spent monthly by each staff role on each patient, as well as necessary startup costs.

Once the chromosome has been parsed, and the relevant patient risk reductions have been applied for each patient we must then find an outcome of the trial. For patients which had no readmission, we retain that value. This does assume that applying an intervention to a patient would not increase their risk. As there were no articles found in which the
trial group had a higher readmission rate than the control group, this is a reasonable assumption. Those patients which had a previous readmission, but are receiving no interventions retained that value. Lastly, patients which had previously been readmitted, but now face a risk reduction, have a chance at preventing readmission. Whether the patient is prevented from having the readmission is simply based on the probability of the readmission reduction. This is assigned randomly at each algorithm run.

At this point in the algorithm, we are aware of all patient outcomes, we are also aware of the number of patients in each intervention group. From this information we calculate the rate of readmission and the excess readmission ratios. From this information we are able to calculate the overall profit or loss based on model implementation. This overall profit or loss is a value weighted as the fitness function of the genetic algorithm.

5.2. Enterprise Economic Concerns

When attempting to find the proper solution, it is necessary to replicate as closely as possible the intricacies of the real world system which we are trying to model. These come in the form of additional constraints which should be imposed on the model. We have already discussed that one of the main components of a program’s cost is the cost of the staff. To some extent, the already available internal staff of an organization might be accomplishing the necessary interventions. However, especially as the number of patient’s increases, it becomes unlikely that there would be such a high number of available staff. In order to fill those staffing roles, additional staff members would need to be hired. This hiring cost should also be included. This is done by utilizing the placement costs for each of the roles as defined by the organization. There are two staff roles, the cardiologist, and the pharmacist which do not require this additional cost. One
of the ways that we have avoided this is by limiting the amount of labor which can be provided by these roles. In reading the descriptions of work which would be accomplished by the cardiologist or pharmacist in the proposed plans, it takes the form of office visits, or in additional discharge work. In both of these instances, it is not reasonable to justify hiring an entire additional staff member.

We also have to consider the limitations which can be imposed upon staff member hiring. It is unlikely to be able to hire qualified employees at in excess quantity. So for the roles of nurse, social worker, case manager we have imposed a constraint by limiting the max number of nurses which can be employed at six, the max number of social workers at 12, and the max number of case managers at eight. These are the total staff roles which can be filled and is calculated from a summation of labor necessary to accomplish all program types.

When discussing how we calculate the costs associated with labor, we need to also replicate how that labor is purchased. When a team member is added, they are added as an employee to the organization. We need to examine not only the amount of labor which is done, but rather to focus on the number of employees which would need to be hired. We treat the capacity of an employee as that of working 40 hours. If additional work is to be filled, then there must be an employee with the available capacity to accomplish that work. If there are no additional employees, and an entire new employee must be added. Labor calculations are conducted based on the number of employees which are needed, rather than on the hourly rate. An exception is made for physicians and pharmacists for the reason specified above.
There should also be a limitation on the amount of capital an organization is able to borrow. Regardless of the return on investment, it may not be possible for an organization to come up with the necessary startup funding for the readmission reduction strategy. We define a few constraints with respect to the overall level of funds. A hard constraint is placed on the overall spending for the readmission reduction strategy. Any solution, regardless of merit in all other respects, which goes beyond the level of this spending should not be considered. An additional borrowing cost is placed on top of spending greater than some level. While an organization may have the available resources to implement a program, it may also be necessary to fund the program beyond this. This serves to incentivize the program creation towards those which are manageable by an organization. While it is still possible to spend an additional amount for a better return, we impose this additional cost or that type of behavior.

In truly defining the optimal readmission reduction intervention strategy for a given organization, we need to define a multitude of organization level constraints as well as those used for calculating costs and returns. It is not reasonable, for instance, for an organization to have an unlimited capital available to it for applying intervention reduction strategies, regardless of whether a return on investment might be expected. There are going to be some organization level constraints which must be imposed on the cost of hiring additional employees responsible for conducting intervention strategies, representations of costs associated with team member intervention activities, available staff time, available physical space, capital goods such as computers, telephones, office space, cost of capital. This should also be extended to cover investments, necessary loans, the cost of different types of staff members', their salaries. Also, hiring should be
limited to either full or part-time with the cost associated with that labor being constant. If somebody is hired on part-time and you only have 20 minutes of work for them to do, then they must be paid for 40 hours of work each week, even though part of it is utilized.

5.3. Results of Experimentation

In the experiment, the full population of CHF patients is represented by a sample of 300 patients. The actual number of heart failure patients discharged monthly averages 70 patients. Our fitness function provides the amount of loss experienced in one month. We have some baselines which will assist us in understanding the outcomes of the strategy.

5.3.1. Base-Line Measures

One possible option is to impose no intervention strategies. Under this circumstance, the losses are based on the readmission penalties placed by CMS, and the losses due to denied charges for our readmitted patients. The maximum penalty allowed in 2015 is 3%. Regardless of the readmission rates, no penalty can exceed 3%. Recall that this reduction is posed on all CMS DRG payments. In this organization, our CMS DRG payments are $30.2 million. The maximum penalty would be $906,000 monthly.

The amount lost due to readmission denials on heart failure patients can be calculated based on the following information. The number of heart failure patients monthly is 70.

The average case charges are $26,000. The baseline readmission rate is 24%. The readmission denial losses average $436,800 monthly. The total possible losses for no intervention reduction programs would average $1.32 million monthly. For this organization, the readmission rates for AMI, Pneumonia, COPD, TKA/THA are all below the level of expected readmission rate. The expected rate of readmission for heart
failure is 22.5%, and the baseline for the organization is 24.6%. This alters the maximum losses based on readmission penalties to a maximum of 0.5%, or $151,000 monthly. This brings the total exposure to $587,800 monthly.

The goal of the heuristic is to determine the correct assignments of interventions on patients. The total cost of the intervention programs, readmission denials, and readmission penalties must be less than the baseline state of $587,800 monthly.

5.3.2. Results Interpretation

The final solution which was determined for their readmission reduction strategy is defined below.

Chromosome = [0, 0, 0, 0, 8, 8, 139, 0, 0, 36, 0, 0, 139, 50, 50, 0]

The fitness function was returned as 10.49. The fitness function should be interpreted as a multiple of $40,000. As such, this fitness function results in a cost of the organization of $419,600 per month. By utilizing the strategy which is shown, we would be applying the interventions described in program #3 to 8 patients, beginning with those at a risk score of. As well, we would also be applying the interventions described in program #5 to 36 patients, beginning with those at a risk score of 100%. These patients would be receiving visits to nurse led heart failure clinics [81].

Finally, we apply program #7 to 50 patients, beginning at the risk score of 76%. The patients categorized under this risk range would be receiving a standardized nurse case management telephone intervention [77]. As the Program #3 is only applied on a very small patient population, trials are run with it removed as shown.
Chromosome = [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 36, 0, 0, 139, 50, 50, 0]

Making this alteration resulted in no significant difference in the cost of the outcome. From the organization viewpoint, the difficulty associated with the program creation and maintenance becomes lessened. Analyses are based on the program under this alteration. While the fitness function shows no difference, we argue that we gain simplicity.

Under this intervention strategy, our CHF readmission rate would be 19.7%. At this rate of return, compared to our baseline strategy which costs $587,800 monthly we can view this expenditure as a savings of $168,200 per month. The total number of staff which would be required is 2 nurses and 3 case managers. At that staffing rate, filling positions with qualified employees should be possible, and the project could be implemented within a reasonable timeline.

One interesting finding of this result, is that program seven is being applied to patients who were in rather moderate risk group. It would seem more likely that program seven should be applied to patients perhaps at the end of the range of program five. By shifting towards this higher risk profile, and not in extending the number of patients being considered, we would expect to see an increased effectiveness, with no additional costs. However under an examination of the program seven shifted to begin at patient 40 i.e. the solution with the chromosome as shown below, testing shows worse performance.

\[
\text{Chromosome} = [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 36, 0, 0, 40, 50, 50, 0]
\]

this result remains counterintuitive. One explanation for this occurrence is that there is a portion of the patient risk profile for the patients in the range from 40 to about 140 in
which are classification accuracy is significantly lower. If the results of prediction on these patients are not as representative, than we would see the performance degradation in the optimization heuristic. As the optimization shows us that patients within a slightly lower range actually pose more of a risk, this appears to be the case. One benefit of this aspect, is that the heuristic can actually identify the appropriate range of patient risks based on the actual outcome of the experimentations.

5.4. Sensitivity Analysis

When utilizing the heuristic for finding a solution, we wish to be assured of three separate things. We want to know that the solution that we have selected as their optimal is actually the best possible solution. Optimization heuristics can sometimes be
unsuccessful, susceptible to finding a local maximum or minimum, rather than the global. Another possible difficulty, is that small changes to the inputs of the program, which sealed large differences and final solutions. As well, we also wish to examine the model convergence to verify that the solution selected is correct. The verification of these points is known as sensitivity analysis. We follow three separate experimental guides. As we look at the genetic algorithm performance, we wish to be assured that the value finally selected by the model was reached by a general consensus of chromosomes within the population. The plot below shows that the initial generation is widely diverse. It includes chromosomes which develop a fitness score near 40. As the model progresses, we can see that the generations tend to become more homogeneous. We can also see that the tendency is towards a low fitness function. In general convergence happens at around generation 150. Due to the variable nature of the model and the inclusion of random parameters it is not odd that the convergence will level at some point.
Figure 21: Showing model performance evaluation as the generation increases
In addressing the possibility of becoming caught in a local optima, we begin each of our scenarios with a completely randomized gene pool. By conducting multiple runs, each run, utilizing a completely randomized initial starting population we can compare the solutions which are found. While it is likely for genetic algorithms to become stuck in local optima, it is unlikely for this to happen repeatedly, and especially not in the same location. If we can run consistent trials which show within reasonable ranges the program inclusion parameters, in this case, the starting position and the amount of patients included, then we can be reasonably assured that the solution is global.

Figure 22: A plot of the difference in the fitness values as the generation increases
Figure 23: Plot which shows the starting point and the length for the vector of application of program 1 to patients.

Figure 24: Plot which shows the starting point and the length for the vector of application of program 2 to patients.

Figure 25: Plot which shows the starting point and the length for the vector of application of program 3 to patients.
Figure 26: Plot which shows the starting point and the length for the vector of application of program 4 to patients.

Figure 27: Plot which shows the starting point and the length for the vector of application of program 5 to patients.

Figure 28: Plot which shows the starting point and the length for the vector of application of program 6 to patients.
From the analyses shown above, we see that there is a general consistency in the strategy defined by multiple runs. There is a much higher variability in the start length, and we can also see outliers in that regard. However, when the length the vector is near zero, the actual outcome of the experiment will not be affected by the start length. A plan which has a start length of zero and the length of zero will score exactly the same as a plan with start length of 100 and the length of zero. By this analysis we can see that programs 1, 2, 3, 4, 6 and 8 were not conducive to a good strategy. By viewing the length across each of

Figure 29: which shows the starting point and the length for the vector of application of program 7 to patients.

Figure 30: which shows the starting point and the length for the vector of application of program 8 to patients.

From the analyses shown above, we see that there is a general consistency in the strategy defined by multiple runs. There is a much higher variability in the start length, and we can also see outliers in that regard. However, when the length the vector is near zero, the actual outcome of the experiment will not be affected by the start length. A plan which has a start length of zero and the length of zero will score exactly the same as a plan with start length of 100 and the length of zero. By this analysis we can see that programs 1, 2, 3, 4, 6 and 8 were not conducive to a good strategy. By viewing the length across each of
the runs, we can see that in general the number of patients was very near zero. Within programs five and program seven, this is not the case. As well there is a fairly consistent trend within these program options with regard to start length and length with respect to the runs conducted.

The other difficulty which we can have with this type of search heuristic is that small changes in input parameters can have large consequences on the optimal solution. This is by far, the more worrying prospect in this specific application. As hard as we may try, there are a fair number of assumptions that we have to make when we are attempting to transfer from literature to specific program utilizations. Some articles actually gave the specific amounts of time which were spent in dealing with patients by each role. However, what is usually provided is a general description of the process. From that general description, we can use our reasonable standard for those types of activities. For instance, we know that a follow-up visit with the cardiologist may take about 30 minutes. Or, we know that a case management type telephone interview with the patient to review symptoms and adherence to proper self-management can be approximately 15 minutes. The same is true with most aspects of program descriptions. Best judgments are used with regard to activities like education follow-up, medication reviews and so on. Given that there will be at least a moderate difference in utilization of each resource; we wish to know that the search heuristic is robust. We wish to know that it still finds the same or near the same optimum solution for the readmission reduction problem, regardless of small changes which may occur in the input parameters.

In order to test this, we utilize experiments which vary the input parameters and assess the changes which occur on the program outcomes. We will conduct this testing at
separate rates. The baseline rate will include the best estimates. An inflated rate will include an additional 10%. As well, the deflated rate will be a reduction of 10%. This will be applied to each sector of input information. As we conduct these trials, we wish to see the genetic algorithm perform to the same standards as is seen in the baseline model. The results of these experiments are shown in the following figures.

In order to test this, we utilize experiments which vary the input parameters and assess the changes which occur on the program outcomes. We will conduct this testing at separate rates. The baseline rate will include the best estimates. An inflated rate will include an additional 10%. As well, the deflated rate will be a reduction of 10%. This will be applied to each sector of input information. The input parameters, we utilize are the monthly labor which is spent on each patient, the resources which are needed as input for program startup, and lastly would be program effectiveness.
Figure 31: The performance under the 'inflated' rate
Figure 32: The performance under the best estimated rate
Figure 33: The performance under the 'deflated' rate
5.5. Comparison to Other Solution Heuristics

When utilizing the heuristic for determining the solution we wish to compare our results to those found by other optimization methods. Our goal is not to prove the best choice for the optimization algorithm, but rather it is to support the decision found in our previous experimental results. Global optimization methods fit into a variety of categories. Exact methods include naïve approaches, exhaustive search, branch and bound algorithms and similar. Naïve approaches include algorithms like uniform grid search, space covering and pure random searches. One limiting factor occurs in higher dimensional problem spaces [86]. Searching for the global optimum by such methods can become too expensive [87]. This prevents their use in this application. Deterministic strategies that search by using the gradient of the function may also be difficult to use in this application due to the non-smoothness in the solution space due to jumps in costs imposed by things like employee hiring and other imposed penalties [88].

Heuristic strategies include evolution strategies, simulated annealing, tabu search or ‘globalized’ extensions of the local search methods [86]. Tabu search could be applied to this type of problem by discretizing the vector solution [89].

While there are a variety of options available to use, we have opted for simulated annealing. Introduced in 1983, it has been used in a wide variety of applications on both discrete and continuous problems [90]. This algorithm operates by selecting random points. New points may be selected even if they raise the cost, a mechanism which allows the algorithm to explore more of the search space. As the temperature of the algorithm lowers, the algorithm becomes more stringent in the solutions it will accept. We use the same solution evaluation algorithm which is defined for the genetic
algorithm. We are interested in comparing the solution strategy, as well as the estimated losses under that strategy. The solution found is listed below.

Solution = [51, 0, 11, 3, 0, 0, 0, 0, 0, 52, 0, 0, 40, 61, 0, 12, 0]

Recall that the solution only indicates using a strategy if the secondary length is non-zero. This allows us to convert the solution to the equivalent form below.

Solution = [0, 0, 0, 0, 0, 0, 0, 0, 0, 52, 0, 0, 40, 61, 0, 0]

This is comparable to the solution found by the genetic algorithm. Moreover, the projected losses which are determined under this strategy would be a loss of $452,000. This too is similar to the projected losses of the genetic algorithm solution. There is an aspect of randomness injected into the algorithm outcome, which preempts algorithms from finding the exact same solutions. This can explain the minor differences which we find within the solutions presented.
CHAPTER VI
FUTURE RESEARCH

While this work shows promise, there are several topics which must be concluded before it can be considered complete. Largely in the investigations presented here detail the overall process. They also show the effectiveness of the process with real-world data. Beyond this we see that the possibilities for approaching the problem in a holistic analytical fashion are valid. In some cases work remains yet to be done, and in others the process utilized should be altered to some extent for improved results. Our aim initially was to prove the validity of the topic, rather than to find the optimum solution.

6.1. Method for Applying Approach in Practice

In this work we have shown how we are able to combine predictive modeling with published intervention literature and optimization heuristics to determine the strategy with the best return financially on the readmission reduction problem. However, there are some details for implementation ideas which should be touched on. In order for the strategy to work effectively in a hospital or hospital system, new patients are going to be targeted. In order for this to happen, the risk score of the patient needs to be determined. The models used to determine the patients risk have already been created, so determining the patient risk score will only necessitate collecting the necessary patient information which is used in the predictive model. Much of this information is already routinely collected, however some of this information would require additional discussion with the
patient. Once the risk or is generated, it can be determined which program or programs the patient would be included in.

In order for this process to work efficiently, it could be useful for the learning algorithm to be embedded within the hospital's EMR. Essentially, it would operate as a secondary system, able to populate the patient information by drawing from the information already collected in the EMR. The outputs required would simply be a yes or no for each intervention option.

6.2. Intervention Options

The first area which would be a topic for additional research is in the collection of, and processing of intervention articles. What we present in this proposal is the method which would be utilized in condensing down articles. While the collection of articles is simple, the processing of the articles would be time-consuming, although certainly possible. However we have not approach this topic with as much rigor as could be applied. Future research could detail multistage reusable framework for condensing literature on trials for interventions. It would cover the aspects related to the acquisition of articles such as which sources should be covered, and the search methodologies which should be employed. It could also discuss how these articles could be grouped for in group and outgroup comparisons. It could also detail more specifically how estimates could be created for plan details such as staff involvements, and ancillary costs related to applying the interventions.

Another aspect of the intervention review would be to define formally the method which would be used for taking results obtained on a certain population and transforming those into those obtained or would be expected to obtain from the hospital standpoint. In
order to accomplish this, the goal would be to simply utilize the hospitals' own data to define its' representative population and to use a modifier, which would transform the expected results based on that population. For instance, if an organization presented in article which showed a 10% reduction in readmissions within the patient population specified, rather than utilizing that organization's quoted readmission rate in that group as the output, it would be more likely and more understandable to utilize the percent reduction which they experience. In this way were able to shift or adjust for differences that might occur between hospitals.

6.3. Sampling

Additional to the feature selection, the oversampling methods should also be reconsidered. There are several options which are available to us in the accomplishment of oversampling. In this area too, we may also wish to consider multiple options attempt and implementation of each compare the affected results, and then make a final determination of which oversampling method to employ. If we simply limit ourselves to one method, we can have no sense of certainty that the method we selected gives the best results. As much of the problem relies on the accuracy of the readmission risk prediction, it is crucial that the results found in the model created utilize every opportunity for improvement. The risk prediction is the beginning of the pipeline in our analytical framework. As such, if the results are tainted than the idiom garbage in garbage out certainly applies. Moreover, if the results of the predictions are poor, than the ability to find an optimal strategy becomes much more difficult. It is simply harder for the heuristic to define an optimal solution. Beyond simply altering the oversampling method, we should also consider the addition of more patient data.
Much work has gone into the comparison of different learning algorithms. Often the comparison attempts to draw the conclusion that one algorithm is superior to others, as in a work by Banko and Brill[91]. They examine the effective sample size or training example size on overall model accuracy. They utilize for state-of-the-art learning algorithms in a comparison. As can be seen from the figure above, any of the algorithms can perform and outperform any of the others. Another interesting trait of these learning
algorithms is that every single one performs better given a larger example size.. This can give some large credence to expanding the sample size utilized in this study. While in this instance, the machine learning task is focused on natural language text disambiguation as well as confusion set disambiguation, it is not unreasonable to assume that our specific learning task would function in the same form. In scenarios where a larger data set is used, results can often be improved by inferior methods simply due to the addition of more training and testing data. As the data set we use here is a relatively small the addition of more patient information would certainly pose a large benefit.

6.4. Missing Data Imputation

in the initial description of the data, I point out that we begin with an initial 652 patients in our sample set. However for all large portion of these patients, there is no collected case management admission interview. Only 452 patients have the entirety of information. In the analysis we handle the simply by moving those patients which do not have this information. However it is important to note that there are number of methods available to us in handling this type of situation. There are really three possible solution to this problem. We can either drop patients with missing information, drop the fields with missing information and keep all patients, or we can attempt to fill in the missing information. Each of these pose some drawback and some benefit. Each of these will be discussed in turn.

6.4.1. Drop Features

one possible method for dealing with these missing values would be to remove those features which contain missing values. In this case, that would mean that all the information contained in that care coordination admission assessment would be removed.
The obvious drawback of this type of strategy for dealing with missing data is that any of the information which could be useful in determining the readmission which is contained in these features is now gone. Unfortunately, as we have seen in the feature selection results, much of this information was useful in the prediction of CHF readmission risk. The benefit, would be that we are able to input the stay lab and diagnosis related information from another 200 patients. The question becomes which is the better trade-off.

6.4.2. Drop Patients

the solution we adopted in our analyses was to drop the patients with the missing information. This represents a loss of 200 patients. What are losing, is information that these patients could have provided on the rest of the variables. Essentially there was an additional 50% of patient data which was collected and ready to be analyzed but which is now been removed. While this represents a significant loss, we do retain some benefits. By not including these patients, we are able to keep the care coordination admission information which provides us with the social characteristics about a patient. Also, this information is pure. There are no invented or imputed data within the set. However, it could be possible to create this information without doing adverse damage to our predictive model performance.

6.4.3. Imputation

Rather than removing either the variables or observations which contain missing data, it is possible to fill in or impute those missing values. There are a wide variety of approaches that can be used to accomplish this task. They range from simple to complex. One large benefit of these methods is that they are able to retain the full sample size
which can provide improvements in bias and precision [92]. There are two main aspects which should be considered when choosing imputation techniques. We need to know whether the data we will be imputing is categorical or continuous. We also want to know whether the data is missing completely at random, missing at random or missing not at random. In our case we would categorize this data as missing not at random. In the way that this admission survey is conducted, certain patients which are believed to be at high risk are always selected. This imposes some rationality or dependency between the type of patient and the survey being conducted. Unfortunately from this information we need to understand that there is an imposed bias within the values that are collected.

One method for addressing this type of situation is to use single imputation methods, which include main mode substitution, dummy variable control, or conditional mean substitution. Under mean mode substitution we would be replacing the missing values with the mean or mode from the rest of the data. One disadvantage is that it reduces the variability in the data and also would weaken covariance and correlation estimates due to neglecting the relationships between variables [93]. Another simple approach is the use of a dummy variable, whereby we introduce a variable containing whether the value was missing or not for a given observation. Missing values are imputed to a constant. A benefit is that it allows us to retain the missing value status but still comes with the disadvantages posed by the previous method [94].

Another approach would be to utilize model-based methods such as Maximum Likelihood Estimation, with identifies parameter values that achieve the highest log-likelihood. One advantage is that it uses the full information from complete cases and incomplete cases to calculate the log likelihood. Unfortunately standard errors tend to be
biased downwards [95]. Beyond this, there is also a whole host of other algorithms which can be used to the same effect. The only way to know for sure which approach provides the best predictive model performance is to try a few methods. It should also be important to note that when attempting these methods, it is beneficial to revisit the feature selection before moving to the creation of the predictive model [96]. As each of these methods will transform or impact the overall data set in different ways, that difference will be reflected in the feature selection step.

6.5. Portability of Research Framework to Other Hospitals

This study as conducted has only shown the application of this framework to a single hospital system. However, it is easy to understand that the same process could be completed from start to finish in any hospital system. All hospital systems possess the data on patients which would make it possible to construct a predictive model. While the way that they hold the information in databases may be different, or the specific types of information may be different this poses little barrier to constructing a predictive model. The largest changes to this framework would most likely be in the actual predictive modeling portion. The information which is collected by hospital systems is to some extent similar, however the amount of information which this particular hospital system collects on the social aspects of patients tends to be a bit larger. There could also be the possibility that some of the information would not be capped in easily mineable forms. Also, it has been shown that when comparing modeling between hospital systems, those features which are relevant or useful in the development of the predictive model are not necessarily the same.
In showing the portability to other hospital systems, the framework must remain functional. So in this way, we must be able to find patient data, and create a predictive model. Then we must be able to summarize and condense literature into a comparative form. Lastly, we must be able to create a heuristic which is capable of optimizing the readmission rate of the hospital with respect to the amount of inputs. On all of these points, portability remains possible.

We can handle each of these points in turn. Can we mine patient data? All hospital systems collect and store patient data. There might be slight differences in how information is stored, or to some extent in what type of information is routinely collected. Both of these factors can pose some types of constraints on what information is readily mineable on patients. For instance, if much of the social information about patients is held in free text, then this information must either be manually curedt or must be passed over entirely.

There can also be some constraints posed on the ability to create predictive models based on the number of patients which the hospital system is seeing. If the patient volume is very low, then it might not be possible to immediately create a well functioning predictive model. Now in this case, our limitation is only on the speed at which we could perform or dive into this framework. If the patient volume is low, then it might just take a longer to collect the appropriate amount of patient information. In either case, the type of information which is readily mineable, or based on the amount of patient information which the hospital system currently has, the only impact is placed on the predictive model performance.
The reason which we might be interested in the predictive model performance for the hospital system would be based on a limited return. As predictive model performance degrades, our ability to see a useful outcome in our optimization and selection of readmission reduction strategies becomes worse. When accuracy is poor, we are more likely to have to increase the group size for readmission interventions in order to achieve the same gains. As well, as the group has becomes larger. We also have a reduced efficiency. Our costs associated with the program are going to be spent less efficiently on actually impacting the rate of readmission reduction.

The second portion of the framework was consolidating literature into are relatively comparable form. In this specific portion of the framework, it is really in no way dependent on the hospital system. Largely, in order to implement our readmission reduction strategy the exact same information needs to be known. As well, there are no changes to the types of literature which would possibly be used. The only limitations we could possibly see, would be based on which of these strategies would be plausible to implement within an organization. And largely these changes would be more readily seen in the heuristics based on the cost of creating a program or based on the amount of labor capacity or physical space or resources of that nature which might be available.

Really, most of the changes that we might see are not based on the literature review aspect of it or on the literature, consolidation portion, but are more readily seen on the types of decisions which might be made based on the optimization heuristic. Now it is important to note that the type of optimization or the way that we present this optimization will not change in the slightest. The exact same methodology can be used in the creation or formulation of this optimization. However, the specific values of
constraints should be adapted per each organization. Now, even though there might be changes in the constraints, it should be noted that none of these changes will in any way stop us from applying the exact same framework in another hospital system.

So in our application of this exact framework into another hospital system, there are some slight issues which might arise. The hospital system might not have the appropriate types of patient information stored in readily mineable forms, or might not have the amount of patient data, which would be required. In this case, these do not prevent the application of the framework, but merely degrade model performance and make the application of this framework slightly less efficient. The only other change which we would readily see, would be based on the constraints which are selected in the creation of the optimization heuristic.

6.6. A Question of Root Cause

A possible critique of the approach used in this framework, is that it does not attempt to divine the root cause for readmissions. The idea being that a patient might be readmitted because they did not get their prescriptions refilled or they might not have had transportation to go see their doctor, or possibly might not even have a doctor. The question becomes whether it is more beneficial to apply these types of interventions. Which might be specific to addressing certain root causes. If the problems were because of medications, than a pharmacist intervention group would be the more effective. This could also be considered in a problem where patients which have an equal risk for also have a differing root causes for this risk.
One of the main reasons for avoiding this type of root cause inference is that actually defining or divining what the root cause of the patients’ readmission risk becomes very, very difficult. In this instance, we were using a random forest model which largely functions as a black box model due to the complexity of the number of forests, and also how those forests votes are aggregated together. This makes it much more difficult to say what aspect of the patient's representation in the data set leads for them to be considered at high risk. Also, as this framework is created, there are no limitations posed on the types of predictive models which might be used. In this way, it is very difficult to assume that we would have the ability to work backwards from a patient's risk or to the root cause of that high risk.

As well, one of the benefits of this approach, is that there is no necessity to conducting the one-on-one social work type interviews, which might be necessary for defining whether the patient might be at higher risk or might have other issues which lead to our readmission being likely. By approaching this problem and applying interventions to patients based on a simple risk score, we can greatly improve or expand how were able to group patients into the readmission reduction interventions which are at the sole force of reducing readmissions.

It is also interesting to note that in examining the types of interventions where one-on-one social work type interventions were conducted which would help patients in obtaining medications or seeing their providers on a regular or scheduled basis, we can find that the efficacy of such programs is actually very similar to the efficacy of programs which just simply apply a unilateral type of intervention to a specific group patients. As such, it is a
better use of time and effort in applying intervention strategies to patients, rather than attempting to determine our root cause for risk.

6.7. Differentiation Between Necessary and Unnecessary Readmissions

There is a difference between unnecessary and unnecessary readmission. However under the program put in place by CMS there is no delineation between the two. They are simply looking at the overall rate. By making the general decision guidelines based on the overall rate, they are attempting to avoid the issue of determining necessary or unnecessary readmission status. There are two separate caveats to this ruling. AMI patients which are returning for certain staged procedures are considered planned. And the same day hospital inpatient readmissions for the same condition to the same hospital are considered planned. As this is the general interpretation taken by CMS, we must adopt the same mindset in our own efforts.

6.8. Summary of Results

While there are some small limitations which have been posed on the study, it has remained remarkably true to replicating the scenario of the hospital readmission reduction problem. This work has utilized a very few simplifications. All of the data used in conducting these trials was obtained from a local community hospital system. All the costs, and estimates are based on that of the hospital system. As well, the patient outcomes have been drawn from actual patient encounters. The solution which was found developed significant cost savings for the organization. The solution proposed was fairly modest, and could reliably be implemented within an organization. By taking this type of problem approach, that of utilizing predictive modeling, literature review, and heuristic searches, we are able to accurately define
specific limitations on groups which should be targeted. We are also able to define which
types of programs represent the most attractive for or this organization. By following this
type of methodology we are much more capable of defining the expected program outcomes.

The ideas presented are applicable to nearly all hospital systems. In order to accomplish
this within an organization, it is necessary that they have significant data, allowing the
creation of the predictive model. Literature access should pose no barrier. As well, the
inputs related to hospital costs and other constraints which are largely financial pose no
barrier. One aspect which may be missing is an individual within the organization
capable in utilizing the types of tools and analyses which were performed. However, this
could easily be obtained via contracting.
BIBLIOGRAPHY


