AN ANALYTICS DRIVEN DECISION SUPPORT SYSTEM
TO INVESTIGATE THE RISK OF NON-INDEX
HOSPITAL READMISSION

A Thesis Presented

By

Yujing Yang

to

The Department of Mechanical & Industrial Engineering

in partial fulfillment of the requirements
for the degree of

Master of Science

in the field of

Industrial Engineering

Northeastern University
Boston, Massachusetts

December 2019
ABSTRACT

Improving the quality of healthcare during hospitalization and after discharging can be realized by identification of 30-day unplanned hospital readmission risk. Prior research suggests that a significant proportion of preventable hospital readmission is attributed to non-index hospital readmission. In particular, followed by the implementation of Hospital Readmission Reduction Program (HRRP), non-index hospital readmission has increased although index hospital readmission has shown a decreasing trend. The existing models in prior researches might not capture the underlying association of predictors with non-index readmission and may lack the reliability and practicability when predicting non-index hospital readmission. Therefore, there exists a critical need to proactively predict non-index hospital readmission in an effort to recommend custom designed post-discharge protocols for patients at risk of experiencing readmission to a non-index hospital. To address this challenge, this study introduces a framework to examine the risk of non-index hospital readmission. Leveraging the state of California hospital discharge datasets, this study uses and compares the predictive models of four machine learning algorithms: logistic regression, random forest, decision tree, and gradient boosting, to predict the likelihood of non-index hospital readmission. AUC and recall scores are used to compare model performance. Results show that the logistic regression model outperforms the other tree-based algorithms, in terms of AUC and recall score. The prominent features shown from the results support previous research findings. This study has the potential to be implemented as a decision support system in clinical setting to help identify the risk of non-index hospital readmission, and thus to recommend effective interventions in order to improve healthcare quality.
TABLE OF CONTENT

1. Introduction ........................................................................................................................................1
2. Review of related research ................................................................................................................3
   2.1 Related work in the field of non-index hospital readmission .........................................................4
   2.2 Review of machine learning algorithms used in hospital readmission ........................................4
3. Methodology ....................................................................................................................................6
   3.1 Feature Engineering .....................................................................................................................7
   3.2 Development and Implementation of Machine Learning Algorithms ........................................8
4. Results and Discussion ....................................................................................................................11
   4.1 Comparison of different models in predicting non-index hospital readmission ..........................11
   4.2 Discussion ................................................................................................................................17
5. Conclusion, limitation and Impaction ...............................................................................................19
REFERENCES .....................................................................................................................................21
LIST OF TABLES

Table 1. Performance of different models on test set.................................13
LIST OF FIGURES

Figure 1. Data preparation, development and evaluation of predictive models..............6
Figure 2. Comparison of AUCs of validation set before/after hyperparameter tuning....12
Figure 3. Comparison of performance metrics for different models..........................13
Figure 4. Visualization of decision tree model.........................................................15
Figure 5. Most important features in decision tree....................................................16
1. Introduction

30-day unplanned hospital readmissions are prevalent events and costly to healthcare systems across the world. Studies documenting that at least 25% of Medicare beneficiaries are readmitted to hospitals within 30 days after discharge and unplanned hospital readmission accounts for approximately $17 billion in annual Medicare spending [3]. In view of this situation, much attention has been devoted to hospital readmission for the Medicare population. As an indicator of the difference in quality of care between hospitals, 30-day readmissions have been embraced by public reporting and payment programs in the United States [14]. Therefore, reducing the hospital admission has been given a national health policy priority.

In order to achieve this objective, the Patient Protection and Affordable Care Act (PPACA) accredited the Centers for Medicare and Medicaid Services (CMS) to penalize excessive accidental readmissions. After that, Hospital Readmission reduction Program (HRRP) was created, through quality improvement strategies, such as providing better care coordination and discharge plans, and motivating hospitals to reduce preventable readmissions. CMS measures hospital readmission rates by comparing hospital readmission rates with national averages. An excess readmission rate of more than 1 means that the hospital is performing worse than ordinary hospitals that allow patients with similar conditions to readmission. CMS calculated the penalty according to the excess readmission ratio, not just payments for readmissions. Hospitals whose readmission ratio are no more than one will be not penalized. For hospitals with ratios greater than one, the higher the ratio, the greater the rate of penalty.

One character of this program is that any readmission penalty is only applied against the index hospital where the initial admission occurred. Thus, readmission of patients who were originally discharged from another hospital will not affect their own readmission rate but will increase the original readmission rate. Prior research has shown that there is a causal relationship between HRRP and the increase in non-index readmissions [15]. However, it shows a great decreased trend in the likelihood of index hospital readmission
when comparing with different-hospital readmissions. Previous research has shown evidence that non-index hospital readmissions play an important role in overall readmissions, about 18-29% [12]. Thus, the introduction of HRRP has prominent effect on non-index readmissions. When compared to index hospital readmission, hospital costs of patients admitted to non-index hospitals are higher. Furthermore, prior studies suggest that non-index hospital readmission may lead to a lack of continuity of treatment and critical clinical information related to patients’ health, which can cause repeated diagnostic tests, delayed medical decision making even worse outcomes [3]. Hence, identifying non-index hospital readmission is of importance to both hospital and patient.

A key element of the problem is that hospital lacks tools for evaluating whether a patient is possible to be readmitted to a non-index hospital. Currently, when identifying readmission risk, only predictive models of hospital readmission are used. Previous studies have showed the models and predictors used in general readmissions. However, when giving insight to index hospital readmission and non-index hospital readmission separately, the predictive models should be different, since there might be differences between the importance of predictors among index and non-index hospital readmissions.

This study intends to provide knowledge and tools necessary for improving readmission prevention practice and thus help curb the increasing rate of non-index hospital readmission. The essential idea is to develop models to predict the likelihood of non-index hospital readmission, and subsequently identify the patient-level and hospital-levels factors related to non-index hospital readmission. The motivation behind this is that research leveraging advanced analytics to proactively predict the non-index readmission is still in early stage. This study is anticipated to be implemented as a decision support system in clinical setting, aiming to help identify the risk of non-index hospital readmissions, which can recommend effective interventions to improve healthcare quality.

The key contributions of this study are:
i. In general, the increasing number of hospital readmissions shortly after discharge is seen as an indicator of the quality of inpatient care and an important factor in the rise in medical costs. While index hospital readmissions have been decreasing following the HRRP, non-index readmissions are increasing, since hospitals limit opportunities of readmission to avoid penalties. To address this issue, this study proactively predicts non-index hospital readmission in an effort to recommend custom designed post-charge protocols for patients at risk of experiencing readmission to a non-index hospital. Identifying those patients and the associated characteristics can help hospitals take corresponding treatment plan in advance.

ii. This study demonstrates that machine learning algorithms superiorly enhance interpretability of the predictors and prediction process within the models. Along with higher predictive power we identify risk factors associated with non-index hospital readmissions that will add to the knowledge base of the existing literatures.

The rest of the paper is organized as follows: Section 2 summaries previous research that relate to non-index hospital readmissions and also presents the state-of-the-art machine learning algorithms used for hospital readmission prediction. Section 3 outlines and describes the proposed predictive analytics framework. Section 4 demonstrates the results and provides a discussion of the findings. Finally, Section 5 concludes with a summary of the work conducted in this research along with some limitations and future research expectations.

2. Review of related research

This section provides a summary of previous research that is related to non-index hospital readmissions. In addition, a brief overview of state-of-the-art data mining and machine learning algorithms that were used for outcome prediction in several healthcare sectors is introduced.
2.1 Related work in the field of non-index hospital readmission

A great many of researches demonstrate the importance of study on non-index readmissions. Prevalence of 30-day non-index hospital readmission has raised much attention, since non-index readmission are common and associated with worse outcomes. Patients who are readmitted to non-index hospitals experience longer hospital length of hospital stay [10]. Previous studies suggest that readmission after surgery is associated with increased mortality [8,9]. Also, in terms of hospital, non-index hospital can result in healthcare utilization, which generates healthcare costs incurring [8].

Previous researches also certify the necessity of study on the non-index hospital readmission. Most of the current state-of-the-art of relevant researches are focused on index hospital readmission. However, there exists difference between the outcomes of index hospital readmission and non-index hospital readmission. Researchers indicate that for patients readmitted to a non-index hospital, the length of hospital stay has increased [3]. In addition, besides the statistical difference between non-index hospital readmission and index hospital readmission. Compared to patients who are readmitted to index hospital, those readmitted to non-index hospitals had small but significant differences in sex, deprivation quintile, comorbidity index and ethnicity [9]. Therefore, researchers suggest that the subject of non-index hospital readmission lack of empirical investigation [12].

2.2 Review of machine learning algorithms used in hospital readmission

A great amount of prior research demonstrates the successful application of machine learning techniques for developing predictive models on hospital readmission. Such studies include but are not limited to non-index hospital readmission prediction. The researches include several clinical conditions, such as pediatric hospital readmission [11], colorectal resection [9], transcatheter aortic [8], major cancer surgery [6], diabetics [5] and heart failure [2].
Among these researches, multiple machine learning techniques are applied. For instance, previous research uses tree-lasso logistic regression to build a predictive model for readmission risk detection based on electronic health records (EHRs). Tree-Lasso logistic regression, a sparse predictive algorithm, can result in an increase of interpretability of the resulting model [1]. Naïve Bayes algorithm has been proved to improve the accuracy of EHRs-wide predictive model [2]. In addition, deep learning related techniques also are applied in hospital readmission prediction. Prior research builds and tests an artificial neural network model based on TensorFlow library. The research result shows that the neural network model can better resolve the complexity and interdependence of various data fields of electronic medical records. [4,5].

In addition, some researches also include the predictors related to non-index readmission. LACE index is the most common predictors used to make prediction. The Lace Index uses four variables to predict the risk of death or non-selective readmission for medical and surgical patients within 30-day discharge: length of stay(L), acuity of the admission(A), comorbidity of the patient(C) and emergency department use in the duration of 6 months before admission (E) [18]. In the level of hospital, some studies use HOSPITAL score as predictors to make prediction, which contains low hemoglobin at discharge, discharge from an oncology service, low sodium level at discharge, procedure during hospital stay, index admission type urgent or emergent, number of hospital admission during the previous year and length of stay [13]. These variables have shown to make prediction of hospital readmission reasonably and successfully.

The current state-of-the-art of relevant researches have focused on specific cohorts of hospital readmission and investigated a selected set of risk factors for non-index readmission, which is arguably narrow in scope. This leads to the following major limitations of the existing studies:

i. After reviewing the state-of-the-art research, the majority of the related literatures are about hospital readmission, including theoretical research and
quantitively study, but little effort on evaluating the prediction of non-index hospital readmissions.

ii. When reviewing the state-of-the-art of non-index hospital, most of the research concentrate on the comparison to the index hospital readmission. Little research is applied data mining and machine learning algorithms to build predictive models.

3. Methodology

This section will present the proposed framework to predict the likelihood of readmission to a non-index hospital. The main purpose is to study which machine learning algorithm applied to the hospital discharge dataset can better predict unplanned non-index readmissions. The second purpose is to identify the predictors that have the greatest impact on models’ performances. The proposed framework as depicted in Figure 1.

Figure 1. Data preparation, development and evaluation of predictive models
3.1 Feature Engineering

3.1.1 Data Source and Study Population
The original source of data was a hospital discharge dataset from state of California for years 2010 through 2014. The California Office of Statewide Health Planning and Development (OSHPD) provides nonpublic datasets of inpatient data collected from California licensed hospitals. These data provide unique identifiers for patients so that it is possible to determine the detailed record for a patient, and patient-level information for clinical and demographic characteristics. To obtain the information of hospital characteristics, this data sources also include American Hospital Association Annual Hospital Survey, Centers for Medicare and Medicaid Services, and the Area Resource File.

The study population included all patients who had a hospital inpatient stay and were discharged alive from a general hospital in California with an initial diagnosis for acute myocardial infarction (AMI), congestive heart failure (CHF), or pneumonia. While most people qualify for Medicare based on age (65 years of age or older), some qualify due to disability or end stage renal disease. For this study, only Medicare patients who were 65 years of age or older are included. Moreover, this study excluded the discharges of ‘patient transferred to another hospital’ or ‘left against medical advice’, patients whose primary diagnosis of mental disease and disorders or rehabilitation related admission, patients did not reside in the State of California and hospitalizations that occurred before February 1, 2010 and after November 30, 2014, since the data set is confined to 2010 through 2014 and a 30-day period is required for identifying readmissions.

3.1.2 Feature Selection
For the predictive models, the target variable, which is binary variable, is classified based on whether a patient was readmitted to a non-index hospital within 30 days. This study includes information both in patient-level and hospital level. For convenience, the predictors are classified into two groups, numerical predictors and categorical predictors.
Numerical predictors include median household income (mhi), hospital competition within a service area as measured by the Herfindahl Hirschman index (HHI), length of stay in hospital, Charlson combordity index. Categorical predictors are gender, age, Medicare status, hospital location (metro or rural), teaching statues (whether or not a member of the Council of Teaching Hospitals), ownership types (nonprofit, investor-owned, public), hospital size based on size of beds (small, medium and large), primary diagnosis (AMI, CHF, and Pneumonia), emergency department visit.

In order to examine collinear relationship between any two independent variables, this study applied correlation analysis on the variables above. The results show that there isn’t significant correlated relationship among the factors, which means the variables above are independent variables and can be used to build the models.

3.1.3 Handling Class Imbalanced Data
The prevalence of non-index hospital readmission was ~25%. Selection of an appropriate performance metric was critical to evaluate the models’ performance. To achieve better results, this study implemented Synthetic Minority Oversampling Technique (SMOTE), which is a technique to over-sample a minority class (non-index readmission). This technique generated ‘synthetic’ samples of a minority class (non-index readmission) and thus a similar distribution of index and non-index hospital readmission in the training set.

3.2 Development and Implementation of Machine Learning Algorithms
This study investigates and evaluates the performance of proposed machine learning algorithms to classify the hospital readmission destination. This problem is framed as a binary classification problem (i.e., target-variable is: ‘0’ if readmitted to index hospital and ‘1’ if readmitted to non-index hospital).

As previously discussed, there isn’t systematic research on prediction of non-index hospital readmission. Thus, this study intends to adopt multiple predictive modeling approaches to present an analytic pipeline for predicting risk of non-index hospital
readmission. More specifically, we intend to assess the performance of traditional logistic regression technique and also tree-based classifiers (decision tree, random forest, and gradient boosting).

3.2.1 Data Split and Ensuring Robust Predictions
This work randomized and divided the data obtained into three sets: training, validation and test set, in a 70:15:15 ratio. Here, the training set is applied to train the model and tune hyperparameters. The validation set is used to make decisions on how model improved after hyperparameter tuning. The test set is for measuring the generalized performance of different models. As the dataset is imbalanced, SMOTE is performed to balance the training dataset and get a 50/50 distribution of index and non-index class.

Moreover, this study uses cross-validation on the training set to make sure that the models are both scalable and robust and are not over-fitting on the training set. Also, it ensures that the models are trained and evaluated on the same data to ensure the performance comparison across models for the exact same data being trained and evaluated upon.

3.2.2 Basic Models
The choice of models is governed primarily by the aim to make prediction of non-index hospital readmission, along with the understanding of the most important factors. Thus, while model accuracy is important, model interpretability in order to devise corrective measures is a key criterion for model selection. The models implemented include:

Logistic Regression: The outcome of logistic regression is the form of log likelihood; it can help understand the relative influence and statistical significance of each factor on the probability of non-index readmission.

Decision Tree: By iteratively and hierarchically predicting whether someone will be readmitted to non-index hospital, it can discover the relative importance of different factors using a more human-like decision making strategy in establishing this decision.
Random Forest: By combining the outcomes of several decision tree models and then doing a majority voting, random forest provides more robust predictive representations than decision tree models.

Gradient Boosting: By creating a group of shallow trees, the gradient boosting model tries to make improvement on the errors of the trained trees.

This study trained different models as mentioned above. 10-fold cross-validation was used when fitting the model on the training set. Cross validation score was used to measure the performance of cross-validation, which is the mean score of the corresponding estimator for each run of the cross validation. Based on the model fitted on the training set, this work fitted the model on validation set and used the score to evaluate.

3.2.3 Hyperparameter Tuning
As a next step, this study optimized the hyperparameters for the models above. One technique for hyperparameter tuning is called grid search, where you can test all possible combinations over a grid of values. However, it is very computationally intensive for the four models with more than 4 or 5 hyperparameters. Therefore, this study uses another option to randomly test a permutation of them, which is called random search.

The performance of hyperparameter tuning is evaluated by area under the curve (AUC) of training set and validation set, respectively. Through comparing the AUC of the validation set before and after hyperparameter tuning, this study can learn how model be improved and make decision of which model should be used in next step.

3.2.4 Evaluating Performance
This study evaluates the performance of different models on test data by using precision, recall, accuracy and AUC. To determine the best estimator, AUC is critical in this study, since AUC is effective to examine the models’ effectiveness of identifying a rare class
from prevalent class. Besides AUC, recall, which measures the actually retrieved class in total amount of relevant instances, is important here since the goal of this study is to identify the patients with potential to be readmitted to a non-index hospital. Therefore, identifying a non-index hospital readmission is more critical than misclassifying an index hospital readmission.

4. Results and Discussion

This section presents a comparative analysis of four predictive modeling approaches based on some performance measures. The influential features which are significantly important for predicting non-index hospital readmission are also described.

4.1 Comparison of different models in predicting non-index hospital readmission

As mentioned above, this study uses the comparison of AUCs of the validation set to learn how the models improved by hyperparameter tuning. The change of the model AUC value before and after hyperparameter tuning is shown in Figure 2. AUC is very effective in indicating model’s effectiveness to identify a rare class from the prevalent one. So, here we use AUC to evaluate the models’ performance.
Based on the comparative performance of AUC, we can observe that the some AUCs for tree-based models basically outperform logistic regression model, especially the models after hyperparameter tuning. Specifically, the AUC in case of validation set are highest for random forest and gradient boosting models, and it is slightly lower for logistic regression and decision tree. In addition, while the logistic regression model is not improved by tuning, tree-based models do get improved by tuning hyperparameter.

For each model, accuracy, precision, recall, AUC for the test set are presented in Table 1, and a graphic comparison is shown in Figure 3.
Table 1. Performance of different models on test set

<table>
<thead>
<tr>
<th>Model</th>
<th>accuracy</th>
<th>precision</th>
<th>recall</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td>0.61</td>
<td>0.34</td>
<td>0.57</td>
<td>0.60</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>0.70</td>
<td>0.35</td>
<td>0.25</td>
<td>0.55</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.73</td>
<td>0.43</td>
<td>0.18</td>
<td>0.55</td>
</tr>
<tr>
<td>Gradient Boosting</td>
<td>0.75</td>
<td>0.52</td>
<td>0.10</td>
<td>0.53</td>
</tr>
</tbody>
</table>

Figure 3. Comparison of performance metrics for different models

As mentioned in previous section, the prevalence of non-index hospital readmission is ~25%, which takes up a smaller proportion. Choosing an appropriate evaluation metric is critical for selecting the best predictive model. Therefore, accuracy alone cannot be considered as a reliable performance measure. Thus, optimizing the model’s recall is significantly considered in this study, meaning that we expect to identify as many patients as possible who with potential to be readmitted to non-index hospital. Identifying a non-index hospital readmission is more critical than misclassifying an index hospital readmission. In terms of AUC and recall, the logistic regression model outperforms other models.

4.1.1 Logistic regression
The logistic regression model does not perform well at first. The hyperparameter tuning do not improve the model, which indicates that the parameters after tuning is the same as the default setting. After generalizing the model on test set, the accuracy is 0.61, the precision is 0.34, the recall is 0.57 and the AUC is 0.60.

Furthermore, this study obtains some meaningful insights from the model by understanding the Odds Ratio (OR) of features. The top features that have the highest OR are AMI, Non-Profit, Investor, Small-bed, Medi-bed. Odds Ratio which is greater than 1 means the higher likelihood of readmitting to non-index hospital. For instance, the OR of AMI is 1.89, which implies that there exists a positive relationship between AMI and non-index hospital readmission. Compared to non-AMI patients, AMI patients are 1.89 times more likely to be readmitted to non-index hospital.

4.1.2 Decision tree
The decision tree model does not perform well in the training set and validation set. Its AUCs for both are lower than other models. By hyperparameter tuning, the AUC for the training set decreases while the AUC for the validation set increases, which means the hyperparameter tuning can reduce the effect of over-fitting in this model. After generalizing the model on the test set, the results are not good as the logistic regression model in terms of recall and AUC. Its recall score is 0.25, and AUC is 0.55.

As an interpretable tool for prediction, the decision tree model can provide insights of decision process. Figure 4 can tell that the first feature used in deciding whether a patient will get readmitted to a non-index hospital or not, is if the HHI is less than or equal to 0.18. Based on the result, it either follows the true or the false path. If it is true, next level of the tree use AMI to decide the readmission destination of a patient. On the other hand, whether HHI is less than or equal to 0.41 will be used to decide the readmission destination.
Figure 4. Visualization of decision tree model

Feature importance can easily summarize the importance information in decision tree model. From the feature importance plot of decision tree model, it can be seen that HHI, median household income, length of stay in hospital, age and several other features are most important predictors of non-index hospital readmission, especially HHI. These are relative numbers, calculated by comparing the performance of the model with or without given predictors. And the relative scores reflect the importance of different predictors. For instance, if removing HHI from the predictors, the model would decrease predictive power.
4.1.3 Random forest

It can be seen that the process of random forest is similar to decision tree. Initially it overfits the training data, but after hyperparameter, it performs well on training as well as validation set. The AUC score for validation set is 0.595 and improved to 0.633 after tuning the hyperparameter. Also, the generalized results on test set is not good, especially recall, which is 0.18. As mentioned in previous sections, this study is trying to identify the patients readmitted to non-index hospitals, thus recall should be considered as priority. In this view, the random forest model is not good for this prediction.
Figure 6. Most important features in random forest

This model gives more importance to following features: median household income, gender, HHI, Gender, length of stay, age. Most of them are consistent with other two models.

4.1.4 Gradient Boosting
The gradient boosting model does not perform well in terms of recall. When fitting the gradient boosting model along with the 10-fold cross validation, the cross-validation score can reach 81.26%. After hyperparameter tuning, the AUC score for validation set is improved from 0.636 to 0.658. Applying the improved model on test set, the performance of the generalized model is not good as that on training and validation set in terms of recall, which is 0.10.

4.2 Discussion

Leveraging several machine learning algorithms, this study builds a system to identify non-index hospital readmission risks. Unlike many existing researches on predicting hospital readmission, this study specifies the readmission to non-index readmission, which is critical to enrich current research and help reduce non-index readmission.
This study introduces four machine learning algorithms and then comparing the performance of different models, by using accuracy, precision, recall and AUC. Considering the accuracy of predicting outcomes, the tree-based models generally perform well than the logistic regression model. This may be caused by decision boundaries are non-linear or there are complex time-sequence dependent interactions between the things happen to a patient during the hospital stay. However, based on the comparison of AUC and recall, this study shows that logistic regression model, in general outperforms tree-based models. It indicates that logistic regression model can identify as more patients as possible who are more likely to be recognized to get readmitted to non-index hospitals. In consideration of the objective, which is to identify non-index hospital readmission, this study mainly adopts AUC and recall measuring the generalized performance. Therefore, logistic regression performs well. Even so, there exists the need to analyze tree-based models. Out of all three tree-based models, decision tree superiorly performs in classifying non-index hospital readmission and index hospital readmission. As expected, the decision tree model has advantages in transparency and interpretability when making classification. By enabling hospitals to identify critical factors to non-index readmission, and then recognizing the patients who are likely to be readmitted to non-index hospitals, decision tree model can help design the prevention scheme based on the risk levels. Thus, decision tree model also can serve as an ideal application for addressing the challenge of reducing non-index hospital readmission.

This study identified the most important features including patient-level and hospital-level characteristics from each model when predicting the likelihood of non-index hospital readmission. The results from this study show that certain predictors are associated with non-index hospital readmissions. Different models determine difference in importance of predictors. In logistic regression model, primary diagnosis (whether the primary diagnosis of a patient is AMI), ownership types (non-profit, investor), and hospital size based on sizes of beds (small-bed, medi-bed) are more important predictors. In tree-based models (decision tree, random forest), median household income, HHI,
gender, Age (whether the age of a patient is no younger than 65), length of stay in hospital and Charlson index are more important when making prediction.

5. Conclusion, limitation and Impaction

In this study, analytics driven decision support system which aims to help reduce non-index hospital readmission is provided, leveraging hospital discharge datasets to investigate the risk of non-index hospital readmission. The proposed framework also intends to help reveal the critical factors associated with non-index hospital readmission. This study particularly focuses on non-index hospital readmission, a critical issue that has been raised much attention in U.S. This is the first attempt to develop a predictive analytics framework of prediction on non-index hospital readmission. This study tackled the imbalanced data utilizing SMOTE and proposed a set of machine learning based predictive models. The obtained result demonstrates logistic regression model outperforms than tree-based models in terms of AUC and recall. However, the decision tree model has better interpretability in decision process. Also, important predictors are shown from the model. Thus, this study enables researchers and healthcare decision makers to utilize machine learning approaches to predict the potential non-index hospital readmission, and provides the inspiration to prevent potential non-index hospital readmission.

There are several limitations of this study. First, this study only considers the condition of readmission, so there exists bias in this research. The data includes non-index hospital readmission and index hospital readmission, but the condition of non-readmission is not included. Moreover, this study uses predictors based on the previous research and conceptual considerations, some potential predictors which can be generated from the datasets may be ignored but still have effect on prediction models. This may affect the accuracy of the models. Another limitation has to do with the selection of four out of the many other predictive algorithms. There are researchers using deep learning techniques to make prediction and have relatively good results. Thus, using other complex models may improve accuracy of prediction. Finally, this research only considers three clinical conditions that were included in the initial enrolled of the HRRP. More clinical
conditions could have been added and there is a need to assess whether the patterns of non-index readmissions in this study apply to other clinical conditions. Future study is expected to improve the predictive effectiveness of the models. First of all, we need to include complete datasets to eliminate bias in this study. Moreover, quality of features depends not only on the volume of the dataset, but also the variety of features. Thus, the diversity of features should be considered. To date, we are aware of not too much studies consider the zip code of patients. By identifying zip code of patients, future studies can address the problem that whether distance to hospital will have effect on non-index hospital readmission. Moreover, such analysis will also take into account other predictive algorithms to investigate and compare their predictive power with current findings.
REFERENCES


