A Predictive Modeling of Tracheostomy Readmissions

Aabha Latkar BSc, Febriany Lete SST, Maysen Pagán BA

Master of Science in Statistical Practice, Boston University.

Abstract

Hospital readmissions following the completion of a procedure poses significant challenges for both patients and hospitals. Being readmitted could increase the risk of complications for patients and introduce states of distress. For hospitals, readmissions present a strain on resources and their reputation. This project aims to develop a model that can predict whether a patient will be readmitted within 30 days of being discharged. After thorough data preprocessing and feature extraction, two models were trained on patient data from 2018 and tested on patient data from 2019 to determine predictive performance: a support vector machine (SVM) and a random forest (RF). Using Matthew's Correlation Coefficient to compare models, the RF model had the best performance with a coefficient of 0.75 while the SVM had a coefficient of 0.61.

1 Introduction

In the medical field of Otolaryngology, preventing hospital readmissions following procedures such as tracheostomies, total laryngectomies, or mastoidectomies is significant both medically for patients as well as financially for healthcare institutions. Medically, avoiding readmissions can benefit patients' well-being as it reduces the possible distress and suffering experienced from complications from new or returning medical conditions. Financially, preventing readmissions is crucial for hospitals who are paid by capitation. Capitation is a payment system that pays hospitals a fixed amount per patient for a prescribed period, therefore incentivizing hospitals to conduct less procedures and treat patients as efficiently as possible. As a result, hospitals paid by capitation incur the costs that are associated with providing care to patients who are readmitted. Knowing if a patient might be at higher risk of a readmission would allow doctors to increase the effectiveness of their initial interventions and promote a smoother recovery process while maintaining their reputation and quality of care. Therefore, developing predictive models that can predict whether a patient will be readmitted is essential for ensuring the efficiency of healthcare.

This project aims to build a model that predicts if a patient who underwent a tracheostomy procedure is going to be readmitted within 30 days of being discharged from the hospital. For those patients who are readmitted within 30 days, this project also analyzes the number of days until they will be readmitted as well as the most common diagnoses that the patients will be readmitted with. The following analysis for this project was conducted on a sample of 3 million observations from the original data.

2 Data Preprocessing

The data for this project was provided to us from the Healthcare Cost and Utilization Project (HCUP) Nationwide Readmissions Database (NRD). This database provides information on admissions and discharges for patients with and without repeat hospital visits within a given year. To address the research questions of this project, we used data from the years of 2018 and 2019, each of which had three data sets. For each year, there was a `core` data set which contained main information such as the age and sex of the patient as well as the diagnoses and procedures the patient received for this given admission. The `severity` data set provides information on the severity of the condition for a given admission and the `hospital` data set contains characteristics of the hospital the patient was admitted to such as bed size.

2.1 Empty Diagnoses and Procedures

The `core` data set contains 40 ICD-10-CM diagnosis code columns representing 40 possible diagnoses a patient could receive at one admission. Similarly, the data includes 25 ICD-10-PCS procedure code columns representing 25 possible procedures a patient could receive at one admission. One patient may not have 40 total diagnoses or 25 total procedures and as a result, many of the cells in these columns are empty. The first step in preparing the data for analysis was replacing these empty cells with `NA`s which will make future filtering and one-hot-encoding steps easier.

2.2 Creating Binary Readmitted Response Column

The focus of this project is predicting readmissions for tracheostomy procedures. Therefore, our next step was to filter the `core` data set to include only those patients who received a tracheostomy during their admission. The ICD-10-PCS codes for tracheostomies all begin with "0B11" with other numbers and letters following for different tracheostomy approaches such as an open approach or percutaneous approach. We applied the filter function to the 25 procedure columns to obtain all patients who had an ICD-10-PCS code that began with "0B11" in any one of the 25 possible columns. Extracting the unique IDs or Visit Links from this filtered set and filtering `core` to get patients only with those Visit Links allowed us to create a new data frame that contained all patients who were admitted for a tracheostomy and all other admissions by that patient. A binary column was then added to this new data frame which was a 1 if the patient received a tracheostomy at the corresponding visit and 0 otherwise. For those tracheostomy admissions, we added the length of stay (LOS) variable to the days to event variable which gives us a sequence of numbers representing the "date" the patient was discharged. This data frame contains all patient admissions both before and after they received the tracheostomy. As a result, for each patient, we then filtered the data frame again to include only the tracheostomy admission and all admissions that occurred after that visit.

This new data frame now allows us to create our binary response variable that is 1 if the patient was readmitted to a hospital within 30 days of being discharged from a tracheostomy procedure. If a patient had no admissions following their tracheostomy admission, the readmitted response variable was 0. If subtracting the days to event value of the next admission after the tracheostomy from the days to event value of the tracheostomy admission results in a value less than or equal to 30, the readmitted response variable was 1. If the difference of the two days to event values was greater than 30, the readmitted response variable was 0.

We then removed all rows that represented hospital admissions after the tracheostomy to obtain a data frame where each row represented a unique patient's tracheostomy admission along with the binary response variable indicating if they were readmitted within 30 days of the procedure.

2.3 Right Censoring

Right censored data is used to describe data where subjects leave the study before an event occurring or the study ends before the event has occurred. In the context of this project, right censoring can occur if a patient received a tracheostomy in December of one year and was readmitted in January of the following year. The data from the NRD is constructed in such a way that it is difficult to identify patients. The patient IDs or Visit Links change from one year to the next for the same patient. As a result, if a patient were to receive a tracheostomy in December of 2018, we would be unable to determine and therefore predict if the patient was readmitted within 30 days since it is unknown if they were readmitted in January of 2019. To solve this issue, we make our window of inclusion for patients between January and November of a given year by filtering out those patients who were admitted for a tracheostomy in December.

2.4 One-Hot-Encoding

When creating the predictive model for readmissions, we wish to conduct comorbidity analysis which will allow us to include the presence of certain diagnoses as predictor variables. For that reason, we will need to one-hot-encode all ICD-10-CM diagnosis codes. For each unique diagnosis code in the 40 diagnosis columns, we added a column for that code which equaled 1 if the patient received the diagnosis on their tracheostomy admission and equaled 0 otherwise.

2.5 Joining Data Sets

The last step in preparing the data for modeling and analysis was to join the three data sets. Each data set had a NRD record identifier variable which gave a unique code to each hospital admission. Therefore, to our cleaned `core` data set, we joined the `severity` and `hospital` data sets by the record identifier.

3 Exploratory Data Analysis

4 Methods and Analysis

In this section, we address and analyze the three objectives of this project.

4.1 Predictive Models

The main objective of this project is to predict whether or not a patient will return to the hospital within 30 days of being discharged from a tracheostomy procedure. This is a classification problem which influences us to apply machine learning classification models. We look at two such models: a support vector machine and a random forest.

4.1.1 Comorbidity Analysis

Before fitting the two models, we first consider what predictors to include. With over 500 possible diagnosis codes each patient can receive, we choose to select a subset of these codes that we believe would have the largest influence on predicting if a patient would be readmitted. To do this, we first observed the proportion of readmitted patients compared to the proportion of non-readmitted patients who received each comorbidity. Here, comorbidity is defined as any diagnosis code that a patient was diagnosed with at the same visit as their tracheostomy procedure. We plotted these proportions on a bar plot in descending order in hopes of being able to compare the proportions between both groups and determine which codes to include in our model based on the largest differences. This plot can be viewed in Figure A2.

However, with such a large number of unique codes, the proportions are quite small and hard to compare. As a result, we then plotted the difference between the proportions of both groups for each comorbidity. In other words, for each comorbidity we plotted the absolute difference between the proportion of readmitted patients who were diagnosed with that comorbidity and the proportion of non-readmitted patients who were diagnosed with that comorbidity. We then chose a cutoff value of 0.015 which returned 40 comorbidities whose difference in proportions was greater than or equal to 0.015. This plot can be viewed in Figure A3. The top two diagnosis codes with the largest difference in proportions for readmitted and non-readmitted patients is Z515 and J9601 which represent encounter for palliative care and acute respiratory failure with hypoxia respectively.

We also included the following variables that we believe to have an association with whether or not a patient is readmitted: age, sex, number of diagnoses, number of procedures, all patient refined diagnosis related group (DRG), and hospital bed size.

4.1.2 Support Vector Machine

The first model we chose to fit is a support vector machine (SVM) which finds a hyperplane to separate our data into two classes (readmitted and non-readmitted) as well as possible while allowing for some violations to this separation to prevent overfitting. The SVM uses what is called a kernel to quantify the similarity between observations. The most popular kernels are the linear kernel, polynomial kernel, and radial kernel with the polynomial and radial kernels allowing for a more flexible decision boundary. Therefore, on our 2018 training data, we fit three SVMs, one for each kernel using the predictors mentioned in Section 4.1.1. We then plotted the ROC curves to evaluate the performance of each classification model on our 2019 testing data. Figure 1 shows that comparing these three kernels, the linear kernel performs the best with an area under the curve (AUC) of 0.756. Therefore, we continue our analysis with the SVM with linear kernel.

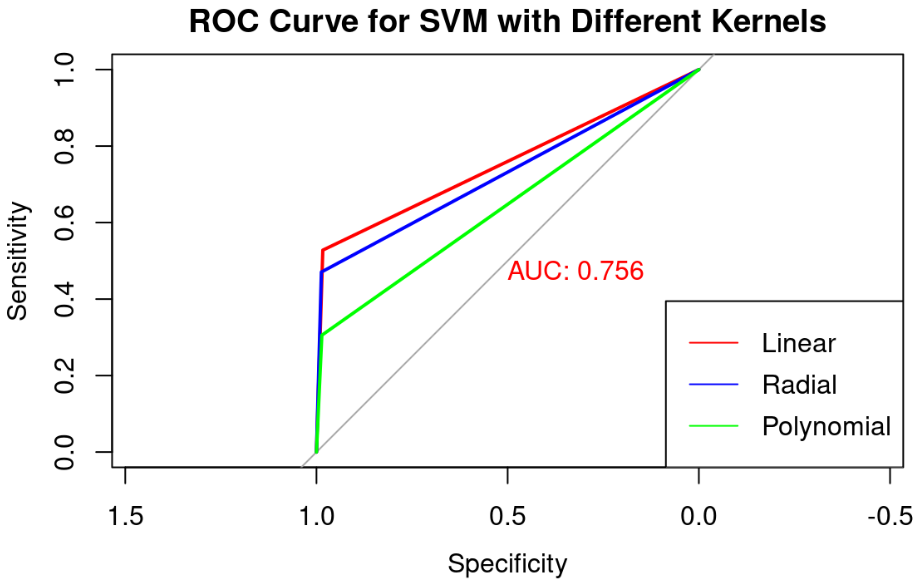
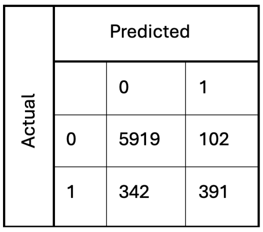


Fig. 1 Plot of the ROC curves for SVMs with linear, polynomial and radial kernels.

Figure 2 below displays the confusion matrix of the SVM with linear kernel model. Due to the imbalanced nature of our data where the readmitted class is larger than the non-readmitted class, the metric of accuracy may be misleading. For example, if 90% of our data belongs to the readmitted class and only 10% belongs to the non-readmitted class, a model that predicts all observations to belong to the readmitted class can achieve 90% accuracy. As a result, to evaluate how well our model correctly predicts observations, we turn to Matthew's Correlation Coefficient (MCC) which takes into account true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). The MCC of our SVM model is 0.6176 and the formula is:

$$MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

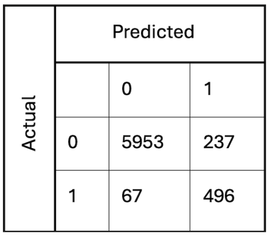


		Predicted	
Actual		0	1
	0	5919	102
	1	342	391

Fig. 2 Confusion matrix for SVM with linear kernel.

4.1.3 Random Forest

The second model we considered was a Random Forest (RF) model which takes our high dimensional data and aggregates decision trees to improve overall predictive performance. Figure 3 displays the confusion matrix for the RF and the MCC can be calculated to get 0.7489.



		Predicted	
Actual		0	1
	0	5953	237
	1	67	496

Fig. 3 Confusion matrix for RF.

We observe that the RF MCC is over 20% larger than the SVM MCC indicating that the RF is the better model in predicting whether or not a patient is readmitted within 30 days of their tracheostomy discharge.

4.1.4 Feature Importance

From our best model, we wish to determine which features contribute the most to make predictions using Shapley values. Each classification prediction can be broken down into a sum of contributions from each input variable of our RF model. These contributions are the Shapley values. We can observe the mean absolute Shapley value for each feature across all the data to determine each feature's overall contribution towards the prediction. The top 10 features can be view in Figure 4

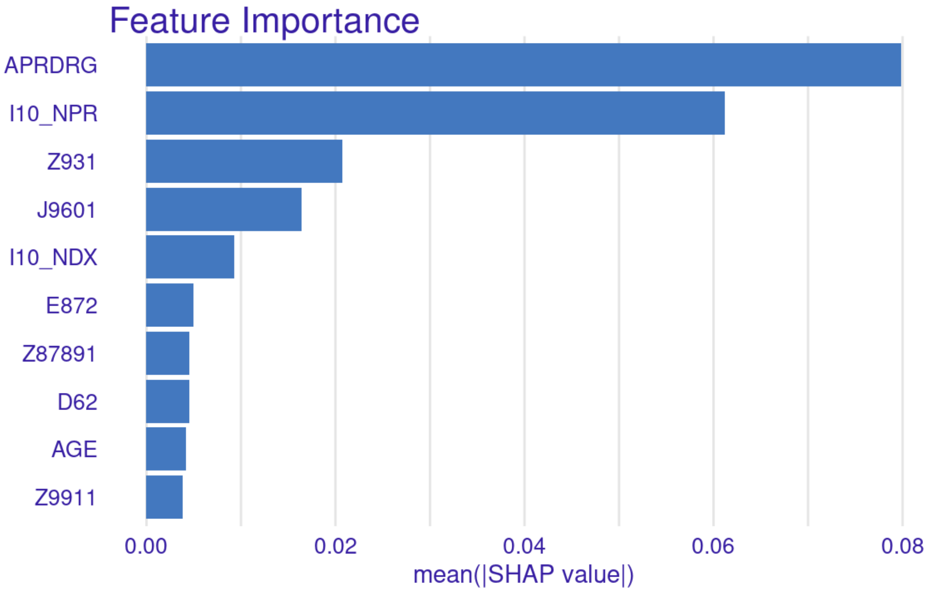


Fig. 4 Feature importance plot of RF model showing the mean absolute Shapley values for each predictor.

Features with higher Shapley values are considered more important as they have a greater impact on the model's predictions. From this figure, the all patient refined DRG marginally contributes the most to the RF's predictions followed by the number of procedures the patient received during their tracheostomy visit as well as whether or not the patient received the diagnosis code Z931 for gastrostomy status.

4.2 Days to Readmission

Another objective of this project was to explore the number of days until patients were readmitted for those who were readmitted within 30 days. Figure 5 plots the counts of days to readmission comparing males and females.

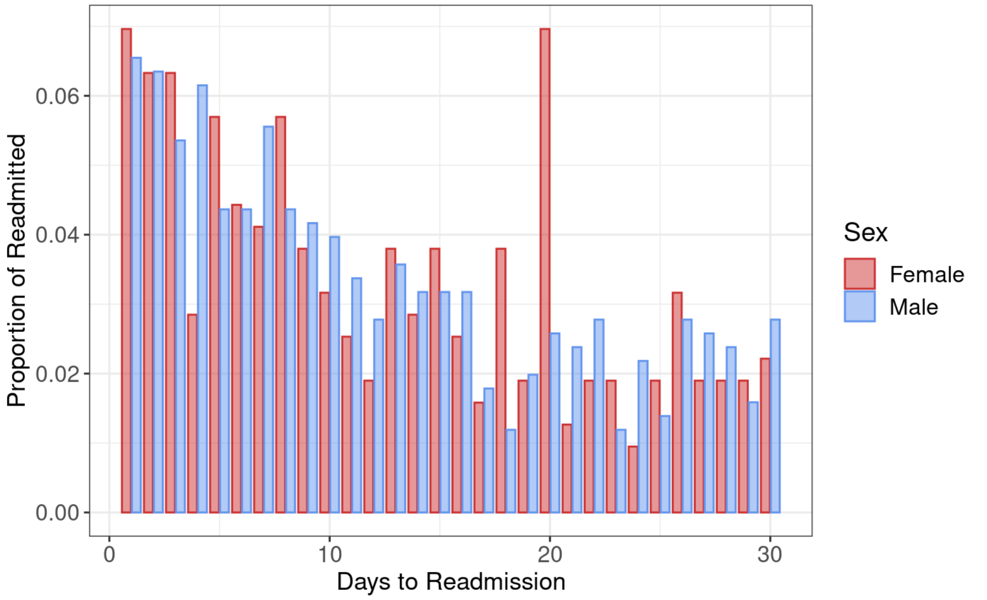


Fig. 5 Barplot of proportion of males and proportion of females who were readmitted to a hospital after a specific number of days following discharge.

From this figure, we can see that most readmitted patients return to the hospital within 10 days of being discharged from the tracheostomy. It is interesting to note that while the largest proportion of males are readmitted after 1 day, there is an equal proportion of females who are readmitted after 1 day as well as after 20 days of being discharged.

4.3 Readmitted Diagnoses

We also created a similar figure looking at the most common diagnoses that patients were readmitted with. For those patients who were readmitted within 30 days of their tracheostomy discharge, we calculated the proportion of males and the proportion of females who received a particular diagnosis code at their readmitted visit. We then plotted these proportions and used the proportion cutoff of 0.15 to observe the top 15 most common readmission diagnoses.

Figure A1 shows that the most common readmitted diagnosis, with an equal proportion of males and females receiving it, is I10 which represents essential (primary) hypertension. This is followed by A419 which is sepsis, a condition in which the body responds improperly to an infection. Similar to hypertension, an equal proportion of males and females are readmitted with sepsis.

The figure also shows that there are 4 common readmission diagnoses that females are diagnosed with but males are not. These codes are F419, E876, F329, and N390 which represent anxiety disorder, hypokalemia, single episode major depressive disorder, and urinary tract infection respectively. Within these top 15 most common readmission diagnoses, there is one code that males are diagnosed with and females are not: E43 representing unspecified severe protein-calorie malnutrition.

5 Conclusion

Acknowledgments. We thank Dr. Anand K. Devaiah, Dean Kennedy, and Dr. Jessica R. Levi from the Boston University Chobanian & Avedisian School of Medicine for providing the objectives of this project, access to the data, and assistance with any questions we had throughout the project. We also thank Professor Masanao Yajima from Boston University MSSP for his guidance, support, and feedback throughout this process.

Appendix A

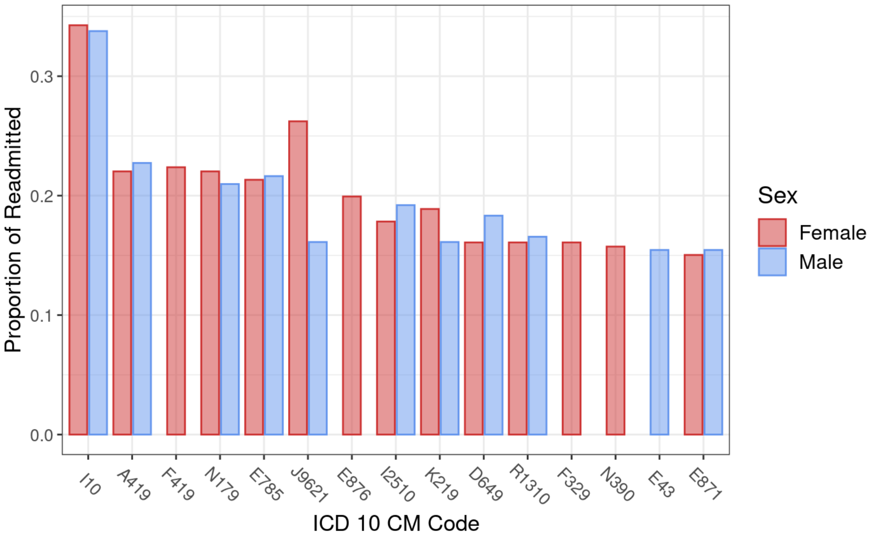


Fig. A1 Barplot of proportion of males and proportion of females who received specific diagnosis codes at their readmission visit.

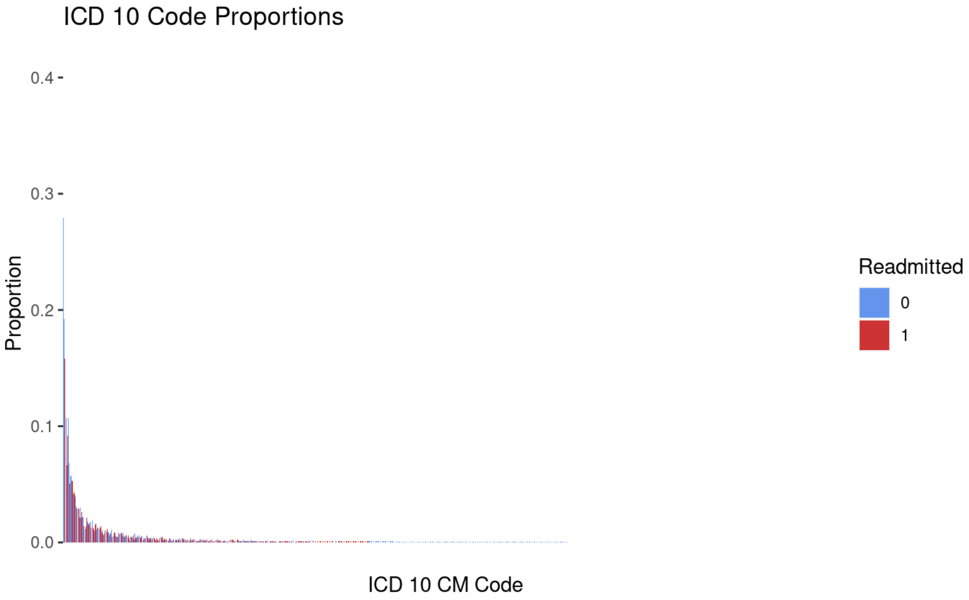


Fig. A2 Barplot of proportion of readmitted patients and proportion of non-readmitted patients who received specific diagnosis codes at their tracheostomy visit.

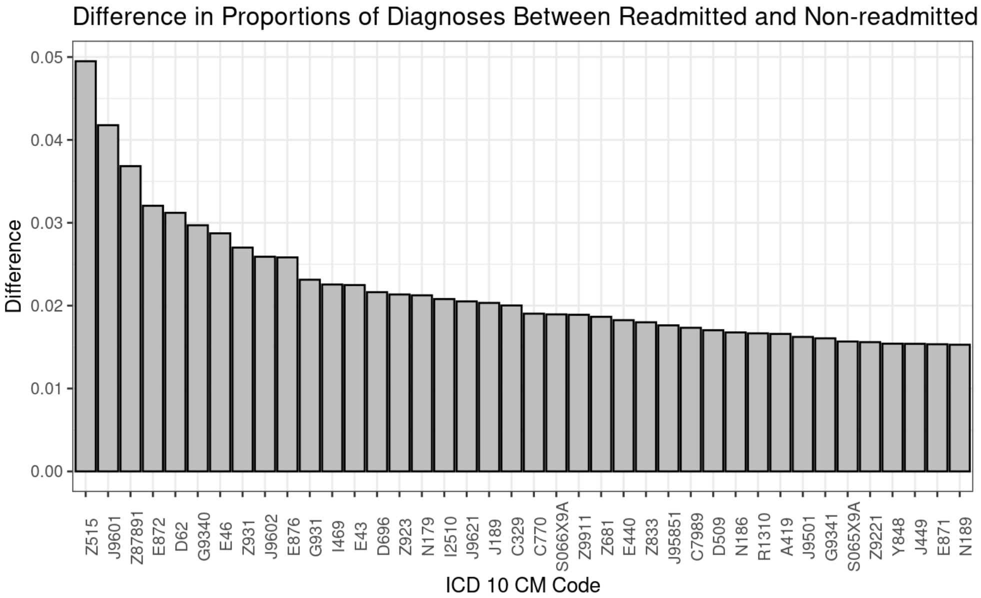


Fig. A3 Barplot of proportion differences between readmitted patients and non-readmitted patients who received specific diagnosis codes at their tracheostomy visit.