Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Xia Lin Date

A Predictive Random Forest Model on Hospital 30-Day Readmission using Electronic Health Records

By

Xia Lin

Master of Science in Public Health

Biostatistics and Bioinformatics

Joel Saltz, M.D., Ph.D.

Committee Chair

A Predictive Random Forest Model on Hospital 30-Day Readmission using Electronic Health Records

By

Xia Lin

Ph.D., University of Notre Dame, 2008

Thesis Committee Chair: Joel Saltz, M.D., Ph.D.

An abstract of

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master of Science in Public Health

in 2012

Abstract

A Predictive Random Forest Model on Hospital 30-Day Readmission

using Electronic Health Records

By Xia Lin

BACKGROUND: Previous studies have employed logistic regression to predict readmission rates and to identify risk factors for readmissions at hospitals. Hospital readmission rates remain high.

OBJECTIVE: To classify patients of 10 diverse subpopulations from Emory hospitals into groups of different 30-day readmission risks using 5-year electronic health records and to validate the applicability of Random Forest on hospital readmission predictions.

METHODS: The information from the 5-year electronic health records at all three Emory hospitals was aggregated into categorical variables and new variables capturing temporal features were also derived. Random Forest algorithms with 10, 50, or 100 trees were used for model construction. Ranking according to the predicted readmission probabilities by the Random Forest models classified patients into groups of different readmission risks.

RESULTS: The risk ranking strategy using Random Forest models successfully separated patients into different risk groups for all 10 subpopulations: cancer, chronic kidney disease, chronic obstructive pulmonary disease, diabetes, heart failure, acute myocardial infarction, pulmonary hypertension, sickle cell anemia, stroke, and history of transplant. Misclassification rates for the top (predicted as "readmitted") and bottom (predicted as "not readmitted") 10% patient subpopulations by risk ranking were also calculated. The models appear to be most effective for stroke patients and least effective for transplant patients. For stroke patients, the readmission rates of patients who are ranked at ≥90%, 75%-90%, 50%-75%, 25%-50%, 10%-25%, and ≤ 10% are 55%, 13%, 11%, 5%, 3%, and 1%, respectively, compared to the baseline readmission rate of 12%. For transplant patients, the readmission rates of patients who are ranked at ≥90%, 75%-90%, 50%-75%, 25%-50%, 10%-25%, and ≤ 10% are 43%, 32%, 24%, 18%, 12%, and 15%, respectively, compared to the baseline readmission rate of 23%.

A Predictive Random Forest Model on Hospital 30-Day Readmission using Electronic Health Records

By

Xia Lin

Ph.D., University of Notre Dame, 2008

Thesis Committee Chair: Joel Saltz, M.D., Ph.D.

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in 2012

Acknowledgements

I am deeply indebted to my advisor, Dr. Joel Saltz, for his guidance and support on my project and manuscript writing process. This thesis would not have been made possible without his guidance and intellectual influence. I also owe huge gratitude to Dr. Jingjing Gao for her in-depth guidance on the project. Her patience, encouragement and support have helped me throughout the whole process from computer programming to manuscript revision. I am also thankful to Drs. Andrew Post and Sharath Cholleti, and Mike Torian for their support to make my work more smoothly.

I am grateful to all my friends in Atlanta who have made my life easier by helping me take care of my daughter at exam times and bringing me lots of joy so I could work on getting my MSPH.

Last but not least, I would like to thank my husband Yang, my lovely daughter Sophia, my mother, my sister, and my parents-in-law for their continuous support. I would not have survived this process without their love.

To Yang and Sophia

Table of Contents

Introduction

Hospital readmissions within 30 days of discharge are a huge burden to both healthcare systems financially and to patients emotionally. About one in five Medicare recipients is readmitted within 30 days, which costs $17 billion in healthcare spending (1). The 30-day readmission rate has been increasingly recognized as an indicator of hospital quality and efficiency of care (2) despite that the evidence of fair comparison (standardized readmission rate) between hospitals has been elusive. Centers for Medicare & Medicaid Services have published risk-adjusted rates for heart failure, pneumonia and heart attacks through the consumers' website, Hospital Compare (http://www.hospitalcompare.hhs.gov). Medicare may reduce payments to hospitals with higher 30-day readmission rates in the near future (3) and as a common practice, private insurance companies may follow Medicare to do the same. Consequently, it is a great incentive for hospitals to make efforts reducing readmission rates.

There are many reasons causing patients to return to hospitals. It may be a new condition, a recurrent exacerbation of a known chronic condition, a complication resulting from previous medical or surgical care, or premature discharge (4). Although it is not easy to predict each specific event, readmitted patients do share certain characteristics. Identification of these factors could benefit both patients by ensuring quality of care and hospitals by saving huge cost. Studies have identified risk factors that are associated with 30-day readmission for general admitted patients and specific patient population (3-8). The general risk factors for readmissions are age, co-morbidities, economic disadvantage, and the number of previous admissions (9). Amarasingham et al. found that the

significant readmission predictors for heart failure patients were marital status, gender, Medicare status, the number of home address changes, history of depression or anxiety, history of confirmed cocaine use, the number of prior inpatient admissions, and presentation time (10). Kim and colleagues identified that higher co-morbidity burden, belonging to racial/ethnic minority groups with public insurance, living in lower-income neighborhoods, and a history of hospitalization in the last 3 months were associated with unscheduled diabetes patients (11). Kansagara's review disclosed that few published models used socioeconomic factors as potential risk factors (12). Models may work better for some patient populations but not for others indicating that patient-level factors play a key role.

Readmission prediction models have been used to facilitate calculation of risk-standardized readmission rates thereby comparing hospital quality of care and to identify risk factors for specific patient population aiming to reduce readmission rate and save huge cost. (For reviews, see (12-15)). Most previous studies on readmission prediction models were based on multivariate logistic regression. Kansagara et al. did a systematic review on risk prediction models for hospital readmission and revealed that there were only 3 models, developed and tested in large European or Australian cohorts, showed c statistics of 0.70 or higher. For the US-based studies, 9 models showed c statistic ranging from 0.55 to 0.65, indicating that the discriminative ability was poor (12).

The wide adoption of electronic health records (EHR) by hospitals has offered great opportunities for retrospective studies. The large databases provide more information

than ever but it is also challenging to extract the most useful information. Traditionally, logistic regression has been used for hospital readmission modeling, as shown in the large body of literature on this topic (5, 10, 16, 17). A large number of models have been built and key risk factors for diseases have been identified (12). Despite the power of logistic regression on binary outcome prediction, it has limitations. Logistic regression makes assumptions such as the independence between observations and that the independent variables are not linear combinations of each other. It may also encounter complete or quarsi-complete separation when the model perfectly or nearly perfectly predicts the response. This problem may be fixed by regrouping of categorical variables and categorizing continuous variables with smaller dataset, however, with larger dataset, it can be challenging to obtain a solution. Recent data mining techniques have been applied in healthcare and medicine (18, 19) although its application on readmission prediction / classification is still a new field to explore.

Data mining is a powerful methodology of exploring large amounts of data to build knowledge and discover unknown patterns or relationships (20, 21) . “Data mining” and “knowledge discovery in databases (KDD)” are sometimes interchangeably used although there are technical differences between them. KDD is “an automatic, exploratory analysis and modeling of large data repositories”. Data mining is “the core of the KDD process, involving the inferring of algorithms that explore the data, develop the model and discover previously unknown patterns” (22).

There are many data mining methods that might be applied to readmission prediction. Classification and Regression Tree (CART) (23), a recursive partitioning method, is one of the commonly used supervised learning algorithms. It is non-parametric so it does not make any assumption about the data distribution. With CART, the prediction models are constructed by recursively partitioning data in a way that minimizes the Gini impurity index of nodes generated at each branch of the tree until all data points are classified into mutually exclusive groups (24). CART is one of the tree-building algorithms that use a set of if-then logical (split) conditions to determine accurate prediction or classification of cases. Tree methods are well suited for data mining purposes as it does not require prior knowledge about the distributions of the variables and the association among the variables. Over-fitting is a common problem with classification trees. If a tree is split for sufficient times, it would be able to predict every single case, however, it does not summarize data thus is of no use predicting cases in a new dataset. As a general rule, we should stop splitting the tree when more splits contribute little to the overall performance of the prediction.

Random Forest (RF) was first proposed by Leo Breinman (25) and is now a trademark of Leo Breinman and Adele Cutler. As the name implies, it uses random bootstrap samples of the original sample to construct classification and regression trees thereby making a "forest". For classification problems, each tree gives a classification ("vote"), and the forest chooses the classification having the most votes as the final result. For regression problems, the average of the values predicted by each trees is used as the estimation of

the final outcome. In random forests, cross-validation is not necessary to get an unbiased estimate of the prediction error. Rather, it is estimated internally during the process.

In this study, we classified patients into different risk groups of readmission within 30 days of discharge based on patients' characteristics using 5-year clinical and administrative data (~250,000 encounters and ~200 attributes) from Emory Hospitals and evaluated the classification accuracy using the data mining classifier, Random Forest.

Methods

Data Description

Prediction models were constructed using EHR data from patients admitted to Emory hospitals including Emory University Hospital, Emory Midtown Hospital, and Wesley Woods Geriatric Hospital, between 4/1/2006 to 3/31/2011.

Outcome Variable

Within each hospital, the admission encounters of each patient were sorted by date and time. For each encounter, the 30-day readmission flag is defined as “yes” if the following admission for the same patient occurs within 30 days of discharge. It is defined as “No” otherwise.

Risk Factors

The variables available in EHR include the following: 1) socio-demographic factors, including age, sex, self-reported race/ethnicity, and insurance status; 2) health condition, including primary diagnosis and secondary diagnosis, as recognized by ICD-9 codes; 3) laboratory values and vital signs such as blood pressure, body-mass index, heart rate, and

platelet count; 4) other factors such as discharge disposition (e.g., discharged to home, skilled nursing facility, or other facilities etc.). Socioeconomic status variables were also available in the EHR system but were not used for this study due to the sensitivity of the protected health information.

Not all original variables retrieved from EHR could be used directly for modeling because the specifications may be in-depth detailed. For instance, the specific values of vital signs do not contribute meaningfully for readmission risk classification. Therefore, based on the primary data, categorical variables were created according to the hierarchies of the diagnosis and procedure codes, thresholds in laboratory test results, medication histories, and longitudinal patterns (e.g., chemotherapy followed by surgery within 180 days) in clinical events.

Higher hierarchies of diagnosis and procedure codes can define disease categories such as uncontrolled diabetes, end-stage renal disease. The continuous vital sign and laboratory test result variables are more meaningful in terms of predictive modeling when classified based on standard medical criteria such as low, normal, and high categories. New variables were derived by considering the medication history and longitudinal patterns. For example, multiple myocardial infarctions (MI) is defined as being diagnosed as MI for more than once during the time period when the data was collected. Similarly, the variable “previous hospitalization (true or false)” was defined as whether the patient has been hospitalized prior to the current encounter. Furthermore, if a patient had multiple

previous readmissions, this patient is also flagged because it is highly associated with patients' disease severity and potentiality to readmission.

Patients

Predictive models based on specific patient subpopulation categorized by disease type is of more value than the models based on entire cohort because the specific patients' characteristics related to certain diseases play important roles on model construction. At Emory hospitals, nine patient subpopulations have been identified by Emory Healthcare discharge reengineering committees in the Emory Enhanced Risk Assessment Tool (ERAT) questionnaire and they are diabetes, heart failure, history of transplant, chronic kidney disease, cancer, chronic obstructive pulmonary disease, acute myocardial infarction, pulmonary hypertension and stroke. In addition to these 9 ERAT categories, the data of sickle cell anemia patients were also analyzed in this study.

Planned readmissions to the hospital were excluded from this analysis as they are not contributive for prediction of preventable readmissions. The exclusion criteria are to exclude encounter pairs from analysis if the second encounter in the pair is a rehabilitation, chemotherapy, radiation therapy, or psychiatry encounter (selected by ICD-9 codes or the location of the encounter) or due to giving birth/delivery because such encounters either are pre-scheduled or inevitable.

Random Forest

Random Forest is an algorithm based on multiple classification or regression trees (25). It uses an ensemble of simple tree predictors, each capable of generating a prediction when presented with a set of independent variable values. During this process, many bootstrap

(sampling with replacement) samples of the same size as the original (herein the training) dataset are drawn. In each of these samples, about two thirds of the observations are chosen one or more times. The remaining one third of the original dataset that are not chosen are called out-of-bag (OOB) for that specific sample. Classification or regression trees are fit for each bootstrap sample. Each fitted tree is then used to predict all OOBs for that tree. A Random Forest consists of an arbitrary number of classification or regression trees, which are used to vote (for classification problem) or to obtain the averaged value (for regression problem) for the final outcome. Using ensemble methods generally leads to improvement in prediction accuracy.

There are basically three variable importance measures used by RF. A naive variable importance measure is to simply count the number of times each variable is selected by all trees. More counting means more importance. Another measure is called “Gini importance”, which is based on Gini index decrease. Gini index is a measure of impurity used by CART, where the split of most Gini decrease is chosen at each node. In Random Forest, adding up all Gini decreases for each individual variable over all trees generates a list of variable importance. The third measure is “permutation accuracy importance" measure. If a predictor variable is associated with the response variable, the original association will be lost after permuting the predictor variable. The difference in prediction accuracy using the predictor variable before and after permuting can indicate the importance of the predictor variable. All three measures are often biased if both categorical and continuous predictor variables exist or the levels of categorical variables differ substantially (26).

In this study, 10, 50, and 100 trees were used for model generation and the performances were compared. We did not use more than 100 trees because doing that would not improve the prediction accuracy of the models. All other parameters were the defaults developed originally by Leo Breiman and Adele Cutler. The R codes are available upon request.

Model evaluation and predictions

Each of the 10 original datasets was randomly split into a training set (80%) and a testing set (20%) for Random Forest modeling. The models constructed from the training sets were then applied on the testing sets for risk ranking and misclassification rate calculation.

By Random Forest models, patients were ranked according to predicted probabilities instead of binary prediction using one cut-off value. By default of many data mining algorithms, a predicted probability of 0.5 is used as a cut-off value for "Yes" or "No" prediction. With a dataset that is unbalanced with more outcomes of "No", we found that when ranking according to the predicted probability the misclassification rates were significantly lowered than the ones with binary prediction using a cut-off value of 0.5. With this ranking strategy, we separated patients with higher and lower readmission risks aiming to provide more information assisting the discharge decisions on each patient at hospitals. We also tested the misclassification rates if predicting that all top 10% ranked patients will come back to the hospital within 30 days while the bottom 10% ranked patients will not come back within 30 days.

Results

Data collection

The EHR data set was de-identified by Emory Healthcare Clinical Data Warehouse. The dataset contained 230276 hospital encounters from 4/1/2006 to 3/31/2011.

Derived variables

Derived variables were defined according to diagnosis and procedure codes, vital signs, and laboratory test results (Table A.1). For example, "uncontrolled diabetes" is defined as "at least one of the following ICD-9 codes, either primary or secondary: 25*.02, 25*.03, 707.1; or HbA1c > 9%".

Descriptive statistics

The characteristics of the patients of all 230276 encounters admitted to three Emory hospitals from 4/1/2006 to 3/31/2011 are summarized in Table 1. Chi-square independence test and t test were done for categorical variables and continuous variables, respectively, to show differences between the patients readmitted and the patients not readmitted. Fisher's exact tests were done instead of chi-square tests when less than 80% of the cells show frequencies > 5. Similar tables for all ten patient subpopulations are listed in Tables A.2 to A.11.

The readmission rates of the ten patient subpopulations are summarized in Table 2. The overall readmission rate for the three Emory hospitals is 14% and the readmission rates of the ten subpopulations are (from high to low): sickle cell anemia (34%), heart failure

(24%), transplant (23%), chronic kidney disease (22%), pulmonary hypertension (21%), cancer (18%), chronic obstructive pulmonary disease (17%), diabetes (17%), and acute myocardial infarction (14%). It is worth noting that these readmission rates are calculated based on encounters, not patients. Also, the total number of encounters is not the sum of the encounters of all subpopulations because some encounters may be tagged with multiple diseases. For instance, one specific encounter could be tagged as both cancer and acute myocardial infarction.

Random Forest ranking and predictions

Each subpopulation was split into one training dataset and one testing dataset at the ratio of 4:1. Random Forest models for predicting the probability of readmission were built using the training sets for all ten patient subpopulations. We used 10, 50, and 100 trees for random forest model construction and we did not change the default parameters of random forest in the R package developed by Leo Breiman and Adele Cutler because we did not find that changing parameters improved the prediction accuracy.

Patients from each subpopulation were ranked/classified according to the predicted probability of readmission by the RF model with 100 trees (Tables 3 and 4). Different random splits of datasets generated similar ranking results (very little variation) so only one set of ranking result using the testing datasets is shown. Also, all ranking results (Tables 3, 4, A.12-A.15) from RF models of 10, 50, 100 trees used the same testing dataset for each subpopulation. The purpose of the ranking is to separate patients with high risk and low risk of readmissions so different considerations can be taken at the discharge stage in hospitals. For patients with higher readmission probabilities, more

caution could be used so informed wise discharge decision can be made at the discharge department of hospitals. Among the cancer patients predicted with various probabilities, the readmission rate for the patients who are ranked larger or equal to 90% is 52%. The readmission rates of patients who are ranked at 75%-90%, 50%-75%, 25%-50%, 10%-25%, and ≤ 10% are 28%, 18%, 11%, 6%, and 4%, respectively. This is much more informative than only knowing the base readmission rate of 18% for these patients because we know that the patients who are ranked above 50% are more likely to be readmitted than the patients who are ranked below 50%. Thus the readmission probabilities of the ranked patients are improved by the RF model than the naive probability of 18% (baseline readmission rate) for all patients. The percentages of the improvement for the cancer patients ranked at ≥ 90%, 75%-90%, 50%-75%, 25%-50%, 10%-25%, and ≤ 10% are 189%, 56%, 0%, 39%, 67%, and 78% over the baseline of 18%. Not surprisingly, there is no improvement for the cancer patients who are ranked at 50%-75% risk of readmission. In the cases of other subpopulations, there is sometimes marginal improvement for this category of ranking (e.g., Chronic Kidney Disease). The similar information for all ten subpopulations is listed in Tables 3 and 4.

The rankings by RF models with 10 and 50 trees are listed in Tables A.12 to A. 15. The model with 10 trees has difficulties separating patients (Tables A.12 and A.13). No patients are under the readmission risk ranking of less than 10% for many subpopulations. When the tree number was increased to 50, the ranking classification got relatively good results (Tables A.14 and A.15). Adding more trees only improved the classification a little better so we stopped at 100 trees (Tables 3 and 4).

We particularly paid more attention to the patients (10% of the subpopulation) whose readmission risks ranked at the highest and the lowest by the RF models for prediction purposes. By our strategy, the patients who are in the top 10% (10% of patients who are predicted at the highest readmission risk) are predicted to be “readmitted” and the patients who are in the bottom 10% are predicted to be “not readmitted”. The misclassification rates and standard deviation over 5-fold cross validation (5 splits of training and testing datasets randomly) are shown in Table 5. For the RF model of 10 trees on cancer patients, the misclassification rate of the top 10% patients is 57% ± 2% (mean misclassification rate ± standard deviation over 5-fold cross validation) and that of the bottom 10% patients is 9% ± 1%. When predicted using the RF model of 50 trees, the misclassification rates for the top and bottom 10% patients are 50% ± 1% and 5% ± 0%. The misclassification rates are 49% ± 1% and 5% ± 0% for the top and bottom 10% patients, respectively, when using the model of 100 trees. The misclassification rates by the model of 10, 50, and 100 trees for the other 9 patient subpopulations are listed in Table 5.

Important variables for generating the RF models are listed in Tables A.1-A.3. These variables were selected by the RF models using the "Gini importance" measure, which describes the total decrease of Gini index (impurity measure) for a particular predictor variable.

Discussion

Reduction on hospital readmission rate is beneficial to both patients and hospitals and is especially an important task for hospitals. This study explored the application of Random Forest, one of the powerful data mining algorithms, on readmission prediction modeling. This is among the few publications of Random Forest on hospital readmission predictions. Based on the predicted readmission probabilities generated from the RF models, the readmission risk classification of 10 patient subpopulations separated patients into different groups. This gives better performance compared with binary prediction ("Yes" if ≥ 50%, "No" if < 50%) if using the predicted probability directly from the model. The classifications are consistent with the actual readmission rates for each risk groups. The readmission rates of the higher risk groups are indeed higher than that of the lower risk groups. This indicates the potential application of the RF models assisting discharge decisions at the hospital discharge departments, in addition to the current factors for discharge decision making.

Random Forest is non-parametric and insensitive to the correlations and collinearity between predictor variables. Unlike CART, RF is also resistant to over-fitting. Compared with logistic regression, RF offers unique advantages for readmission prediction using EHR data. Traditional logistic regression usually does not handle too many (e.g., >100) variables very well. Even if one can use univariate logistic regression to do screening first then apply model selection strategies for multivariate models, it is not an easy task because higher p-values from univariate models do not necessary mean their unimportance and cannot rule out the possibility of the interaction between the particular

variable and other variables. In that case, previous knowledge helps but also prevents from finding new predictors from EHR that are unknown.

In this study, we split each of the 10 original datasets into training sets and testing sets despite that RF does cross validation internally within each model generation process. This is to keep the consistency with current literature on readmission prediction modeling and is also a good practice for comparing the performance of RF with other algorithms using the same data.

It appears that the RF models are most effective at ranking stroke patients and least effective for transplant patients with the datasets we analyzed. For stroke patients, the readmission rates of patients who are ranked by RF model with 100 trees at ≥90%, 75%-90%, 50%-75%, 25%-50%, 10%-25%, and ≤ 10% are 55%, 13%, 11%, 5%, 3%, and 1%, respectively, compared to the baseline readmission rate of 12%. The improvement over baseline is 358%, 8%, 8%, 58%, 75%, 92%, for the ranking categories, from high to low. For transplant patients, the readmission rates of patients who are ranked by RF model with 100 trees at ≥90%, 75%-90%, 50%-75%, 25%-50%, 10%-25%, and ≤ 10% are 43%, 32%, 24%, 18%, 12%, and 15%, respectively, compared to the baseline readmission rate of 23%. The improvement over baseline is 87%, 39%, 4%, 22%, 48%, 35%, for the ranking categories, from high to low (Table 4). This is not unexpected because, as shown in Figure A.3, discharge disposition is a strong predictor for stroke patients but not for transplant patients.

The important variable lists provided by Random Forest need to be used with caution (26). They are based on single classification trees, although meaningful for many applications, are not reliable in situations where both categorical and continuous potential predictor variables are used. Even if all predictor variables are categorical, the Gini importance measure is also biased if the levels of categorical predictor variables differ substantially. The continuous variables and the categorical variables with more levels are often favored by RF. However, if the same set of variables are used for different subpopulations as in the current study; important variables are comparable between the subpopulations, as in the case of above-mentioned comparison between stroke patients and transplant patients.

Geographic and socioeconomic variables are not available to this study due to the "Protected Health Information" nature of these variables. Future study may include these variables and they may contribute to the prediction accuracy of RF models (12).

Also, many co-morbidity indices have been shown to be predictors of readmission and mortality and they are Cumulative Illness Rating Scale (CIRS) (27), Charlson Comorbidity Index (CCI) with or without Deyo modification (28, 29), Chronic Disease Score (CDS) (30), Elixhauser comorbidity measurement (31), Index of Coexistent Diseases (32), Kaplan scale (33), and Geriatric Index of Comorbidity (GIC) (34). Among these, CCI is most widely used for predictive modeling. Zekry and colleagues compared the performance of the 6 indices (all above-mentioned indices except the Elixhauser comorbidity measurement) as predictors of 1-year post-hospital discharge

institutionalization, readmission, and mortality in elderly individuals (35). Their univariate logistic regression result demonstrated that high scores for the CIRS, CCI, and CDS indices were found to be independent predictors for readmission at 12 months of discharge in the elderly patient population. When predicting 1-year mortality with univariate Cox regression, the best predictor was CIRS, followed by GIC, Kaplan scale, ICED, and CCI. Therefore, future studies utilizing some of these comorbidity indices as predictors may improve the performance of the RF models substantially.

Summary

This study generated Random Forest predictive models for classifying patients according to hospital 30-day readmission risks. The EHR datasets of ten patient subpopulations at Emory hospitals were analyzed and classification of patient risk groups was successfully achieved. It demonstrated that RF is a powerful data mining tool on hospital 30-day readmission classification. The 10% of each subpopulation that are ranked with the highest readmission risk is predicted as “readmitted” and the 10% of each subpopulation that are ranked with the lowest readmission risk is predicted as “not readmitted”. By doing this, the respective misclassification rate was calculated. This provides much more useful information to the hospital discharge authorities than merely knowing the baseline readmission rate for the subpopulation.

This work has been approved by the Institutional Review Board (IRB) and the protocol number is IRB00054656.

References

1. Jencks SF, Williams MV, Coleman EA. Rehospitalizations among patients in the Medicare fee-for-service program. N Engl J Med 2009;360(14):1418-28.

2. Balla U, Malnick S, Schattner A. Early readmissions to the department of medicine as a screening tool for monitoring quality of care problems. Medicine (Baltimore) 2008;87(5):294-300.

3. Allaudeen N, Vidyarthi A, Maselli J, Auerbach A. Redefining readmission risk factors for general medicine patients. J Hosp Med 2011;6(2):54-60.

4. Silverstein MD, Qin H, Mercer SQ, Fong J, Haydar Z. Risk factors for 30-day hospital readmission in patients >/=65 years of age. Proc (Bayl Univ Med Cent) 2008;21(4):363-72.

5. Hasan O, Meltzer DO, Shaykevich SA, Bell CM, Kaboli PJ, Auerbach AD, et al. Hospital readmission in general medicine patients: a prediction model. J Gen Intern Med 2010;25(3):211-9.

6. Grant RW, Charlebois ED, Wachter RM. Risk factors for early hospital readmission in patients with AIDS and pneumonia. J Gen Intern Med 1999;14(9):531-6.

7. Corrigan JM, Martin JB. Identification of factors associated with hospital readmission and development of a predictive model. Health Serv Res 1992;27(1):81-101.

8. Reed RL, Pearlman RA, Buchner DM. Risk factors for early unplanned hospital readmission in the elderly. J Gen Intern Med 1991;6(3):223-8.

9. Howell S, Coory M, Martin J, Duckett S. Using routine inpatient data to identify patients at risk of hospital readmission. BMC Health Serv Res 2009;9:96.

10. Amarasingham R, Moore BJ, Tabak YP, Drazner MH, Clark CA, Zhang S, et al. An automated model to identify heart failure patients at risk for 30-day readmission or death using electronic medical record data. Med Care 2010;48(11):981-8.

11. Kim H, Ross JS, Melkus GD, Zhao Z, Boockvar K. Scheduled and unscheduled hospital readmissions among patients with diabetes. Am J Manag Care 2010;16(10):760-7.

12. Kansagara D, Englander H, Salanitro A, Kagen D, Theobald C, Freeman M, et al. Risk prediction models for hospital readmission: a systematic review. JAMA 2011;306(15):1688-98.

13. Bahadori K, FitzGerald JM. Risk factors of hospitalization and readmission of patients with COPD exacerbation--systematic review. Int J Chron Obstruct Pulmon Dis 2007;2(3):241-51.

14. Desai MM, Stauffer BD, Feringa HH, Schreiner GC. Statistical models and patient predictors of readmission for acute myocardial infarction: a systematic review. Circ Cardiovasc Qual Outcomes 2009;2(5):500-7.

15. Ross JS, Mulvey GK, Stauffer B, Patlolla V, Bernheim SM, Keenan PS, et al. Statistical models and patient predictors of readmission for heart failure: a systematic review. Arch Intern Med 2008;168(13):1371-86.

16. Anderson GF, Steinberg EP. Predicting hospital readmissions in the Medicare population. Inquiry 1985;22(3):251-8.

17. Novotny NL, Anderson MA. Prediction of early readmission in medical inpatients using the Probability of Repeated Admission instrument. Nurs Res 2008;57(6):406-15.

18. Yoo I, Alafaireet P, Marinov M, Pena-Hernandez K, Gopidi R, Chang JF, et al. Data Mining in Healthcare and Biomedicine: A Survey of the Literature. J Med Syst 2011.

19. Amalakuhan B, Kiljanic L, Hester M, Cheriyath P, Fischman D. A Prediction Model For COPD Readmissions : Catching Up , Catching Our Breath And Improving A National Problem. Am J Respir Crit Care Med 2011;183:A6436.

20. Bellazzi R, Zupan B. Predictive data mining in clinical medicine: current issues and guidelines. Int J Med Inform 2008;77(2):81-97.

21. Goodwin L, VanDyne M, Lin S, Talbert S. Data mining issues and opportunities for building nursing knowledge. J Biomed Inform 2003;36(4-5):379-88.

22. Maimon O, Rokach L. Data mining and knowledge discovery handbook. 2nd ed. New York: Springer; 2010.

23. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. Monterey, California, USA: Wadsworth, Inc; 1984.

24. Maroco J, Silva D, Rodrigues A, Guerreiro M, Santana I, de Mendonca A. Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. BMC Res Notes 2011;4:299.

25. Breiman L. Random forests. Machine Learning 2001;45(1):5-32.

26. Strobl C, Boulesteix AL, Zeileis A, Hothorn T. Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinformatics 2007;8:25.

27. Parmelee PA, Thuras PD, Katz IR, Lawton MP. Validation of the Cumulative Illness Rating Scale in a geriatric residential population. J Am Geriatr Soc 1995;43(2):130-7.

28. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J Chronic Dis 1987;40(5):373-83.

29. Deyo RA, Cherkin DC, Ciol MA. Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. J Clin Epidemiol 1992;45(6):613-9.

30. Von Korff M, Wagner EH, Saunders K. A chronic disease score from automated pharmacy data. J Clin Epidemiol 1992;45(2):197-203.

31. Southern DA, Quan H, Ghali WA. Comparison of the Elixhauser and Charlson/Deyo methods of comorbidity measurement in administrative data. Med Care 2004;42(4):355-60.

32. Greenfield S, Sullivan L, Dukes KA, Silliman R, D'Agostino R, Kaplan SH. Development and testing of a new measure of case mix for use in office practice. Med Care 1995;33(4 Suppl):AS47-55.

33. Kaplan MH, Feinstein AR. The importance of classifying initial co-morbidity in evaluatin the outcome of diabetes mellitus. J Chronic Dis 1974;27(7-8):387-404.

34. Rozzini R, Frisoni GB, Ferrucci L, Barbisoni P, Sabatini T, Ranieri P, et al. Geriatric Index of Comorbidity: validation and comparison with other measures of comorbidity. Age Ageing 2002;31(4):277-85.

35. Zekry D, Valle BH, Michel JP, Esposito F, Gold G, Krause KH, et al. Prospective comparison of six co-morbidity indices as predictors of 5 years post hospital discharge survival in the elderly. Rejuvenation Res 2010;13(6):675-82.

Tables

Table 1. Entire Cohort Patient Characteristics

Variable	Level	Readmitted n (%) or mean±std	Non-readmitted n (%) or mean±std	Simple test* p-value
Amputation Indicator	Yes	977 (3.1)	3253 (1.6)	<.0001
	No	30817 (96.9)	195229 (98.4)	
Bone Marrow Transplant Indicator	Yes	184 (0.6)	851 (0.4)	0.0002
	No	31610 (99.4)	197631 (99.6)	
Encounter 180 Days Earlier	Yes	15954 (50.2)	48195 (24.3)	<.0001
	No	15840 (49.8)	150287 (75.7)	
Encounter 90 Days Earlier	Yes	13850 (43.6)	38245 (19.3)	<.0001
	No	17944 (56.4)	160237 (80.7)	
End Stage Renal Disease Indicator	Yes	4657 (14.6)	13207 (6.7)	<.0001
	No	27137 (85.4)	185275 (93.3)	
Insurance Flag	Yes	30481 (95.9)	187366 (94.4)	<.0001
	No	1313 (4.1)	11116 (5.6)	
Metastasis Indicator	Yes	2183 (6.9)	9049 (4.6)	<.0001
	No	29611 (93.1)	189433 (95.4)	
Methicillin-resistant Staph Aureus Indicator	Yes	358 (1.1)	1121 (0.6)	<.0001
	No	31436 (98.9)	197361 (99.4)	
Multiple MIs	Yes	1324 (4.2)	2806 (1.4)	<.0001
	No	30470 (95.8)	195676 (98.6)	
Multiple Readmits In the Past	Yes	7065 (22.2)	12997 (6.5)	<.0001
	No	24729 (77.8)	185485 (93.5)	
Obesity Indicator	Yes	9807 (30.8)	59906 (30.2)	0.0168
	No	21987 (69.2)	138576 (69.8)	
Sex	Unknown	0 (0)	5 (0)	<.0001
	Male	14707 (46.3)	84460 (42.6)	
	Female	17087 (53.7)	114017 (57.4)	
Race	White	14078 (44.3)	94353 (47.5)	<.0001
	Other	1299 (4.1)	13397 (6.7)	
	Black	16360 (51.5)	90380 (45.5)	
	Asian	57 (0.2)	352 (0.2)	
Pressure Ulcer Indicator	Yes	1392 (4.4)	4185 (2.1)	<.0001
	No	30402 (95.6)	194297 (97.9)	
Readmit Neutropenia Flag	Yes	683 (2.1)	512 (0.3)	<.0001
	No	31111 (97.9)	197970 (99.7)	
Sickle Cell Anemia Indicator	Yes	745 (2.3)	1441 (0.7)	<.0001
	No	31049 (97.7)	197041 (99.3)	
Sickle Cell Crisis Indicator	Yes	660 (2.1)	1145 (0.6)	<.0001
	No	31134 (97.9)	197337 (99.4)	
Uncontrolled Diabetes Indicator	Yes	2184 (6.9)	9361 (4.7)	<.0001
	No	29610 (93.1)	189121 (95.3)	
Age		56.3 ± 17.5	54.7 ± 18.2	<.0001
ERAT Count		1.2 ± 1.0	0.9 ± 1.0	<.0001
Length of Stay		7.8 ± 9.9	5.5 ± 7.6	<.0001

*Two-sample t tests were performed for continuous variables and chi-square independence tests were performed for categorical variables. When less than 80% cells show frequencies >5, Fisher's exact tests were carried out instead of chi-square tests.

Table 2. Readmission Rate Distribution of Emory Hospitals

Subpopulation	# of Total Readmission	# of Total Encounters	Readmission Rate
Cancer	9300	50391	18%
Chronic Kidney Disease	9230	41901	22%
Chronic Obstructive Pulmonary Disease	3705	21203	17%
Diabetes	9292	55093	17%
Heart Failure	5197	21550	24%
Acute Myocardial Infarction	3082	22403	14%
Pulmonary Hypertension	1258	5973	21%
Sickle Cell Anemia	745	2186	34%
Stroke	842	6858	12%
Transplant	1171	5147	23%
Total	31794	230276	14%

Table 3. Ranking of Readmission Risk by Random Forest (100 trees) Model, part I

Subpopulation	Ranking	# of encounter	# of Readmission	Readmission rate	Improvement over baseline
Cancer	>= 90%	1062	552	52%	189%
	75% - 90%	1501	422	28%	56%
	50% - 75%	2621	461	18%	0%
	25% - 50%	2693	302	11%	39%
	10% - 25%	1358	87	6%	67%
	<= 10%	871	34	4%	78%
	Total	10106	1858	18%	
Chronic Kidney Disease	>= 90%	911	462	51%	132%
	75% - 90%	1217	422	35%	59%
	50% - 75%	2269	557	25%	14%
	25% - 50%	2136	300	14%	36%
	10% - 25%	1111	115	10%	0.55
	<= 10%	761	30	4%	82%
	Total	8405	1886	22%	
Chronic Obstructive Pulmonary Disease	>= 90%	447	223	50%	194%
	75% - 90%	631	166	26%	53%
	50% - 75%	1119	190	17%	0%
	25% - 50%	1058	92	9%	47%
	10% - 25%	610	44	7%	59%
	<= 10%	414	32	8%	53%
	Total	4279	747	17%	
Diabetes	>= 90%	1116	537	48%	182%
	75% - 90%	1782	480	27%	59%
	50% - 75%	2879	444	15%	12%
	25% - 50%	2841	274	10%	41%
	10% - 25%	1564	119	8%	53%
	<= 10%	857	29	3%	82%
	Total	11039	1883	17%	
Heart Failure	>= 90%	460	241	52%	117%
	75% - 90%	687	223	32%	33%
	50% - 75%	1058	275	26%	8%
	25% - 50%	1179	204	17%	29%
	10% - 25%	564	58	10%	58%
	<= 10%	412	28	7%	71%
	Total	4360	1029	24%	

Table 4. Ranking of Readmission Risk by Random Forest (100 trees) Model, part II

Subpopulation	Ranking	# of encounter	# of Readmission	Readmission rate	Improvement over baseline
Acute Myocardial Infarction	>= 90%	482	226	47%	236%
	75% - 90%	655	162	25%	79%
	50% - 75%	1223	159	13%	7%
	25% - 50%	1081	59	5%	64%
	10% - 25%	663	21	3%	79%
	<= 10%	423	3	1%	93%
	Total	4527	630	14%	
Pulmonary Hypertension	>= 90%	124	57	46%	119%
	75% - 90%	180	61	34%	62%
	50% - 75%	283	64	23%	10%
	25% - 50%	291	41	14%	33%
	10% - 25%	176	15	9%	57%
	<= 10%	108	6	6%	71%
	Total	1162	244	21%	
Sickle Cell	>= 90%	45	32	71%	122%
	75% - 90%	61	30	49%	53%
	50% - 75%	106	38	36%	12%
	25% - 50%	109	17	16%	50%
	10% - 25%	54	11	20%	38%
	<= 10%	42	7	17%	47%
	Total	417	135	32%	
Stroke	>= 90%	137	76	55%	358%
	75% - 90%	211	28	13%	8%
	50% - 75%	372	41	11%	8%
	25% - 50%	300	15	5%	58%
	10% - 25%	232	6	3%	75%
	<= 10%	105	1	1%	92%
	Total	1357	167	12%	
Transplant	>= 90%	99	43	43%	87%
	75% - 90%	160	52	32%	39%
	50% - 75%	238	56	24%	4%
	25% - 50%	235	43	18%	22%
	10% - 25%	150	18	12%	48%
	<= 10%	94	14	15%	35%
	Total	976	226	23%	

Table 5. Misclassification Rate by Random Forest with 10, 50, or 100 trees

Subpopulation	Classification	Mean Misclassification Rate ± Standard Deviation (of 5-fold Cross Validation)		
		10 trees	50 trees	100 trees
Cancer	Top 10%	57% ± 2%	50% ± 1%	49% ± 1%
	Bottom 10%	9% ± 1%	5% ± 0%	5% ± 0%
Chronic Kidney Disease	Top 10%	54% ± 1%	47% ± 2%	46% ± 2%
	Bottom 10%	11% ± 1%	7% ± 1%	6% ± 1%
Chronic Obstructive Pulmonary Disease	Top 10%	59% ± 2%	54% ± 2%	52% ± 2%
	Bottom 10%	9% ± 1%	7% ± 1%	6% ± 1%
Diabetes	Top 10%	60% ± 2%	53% ± 1%	53% ± 1%
	Bottom 10%	8% ± 1%	5% ± 1%	5% ± 1%
Heart Failure	Top 10%	52% ± 2%	44% ± 1%	43% ± 1%
	Bottom 10%	12% ± 2%	9% ± 2%	8% ± 1%
Acute Myocardial Infarction	Top 10%	61% ± 2%	54% ± 3%	54% ± 3%
	Bottom 10%	6% ± 2%	2% ± 0%	2% ± 1%
Pulmonary Hypertension	Top 10%	57% ± 5%	52% ± 4%	49% ± 5%
	Bottom 10%	13% ± 2%	7% ± 2%	6% ± 2%
Sickle Cell	Top 10%	34% ± 6%	29% ± 5%	27% ± 4%
	Bottom 10%	17% ± 8%	18% ± 7%	16% ± 4%
Stroke	Top 10%	57% ± 6%	50% ± 5%	50% ± 5%
	Bottom 10%	6% ± 2%	2% ± 2%	2% ± 1%
Transplant	Top 10%	56% ± 5%	50% ± 5%	48% ± 6%
	Bottom 10%	17% ± 2%	12% ± 4%	11% ± 3%

Appendix

Table A.1 Derived variables used in readmissions analyses (all definitions are for inpatient encounter data)

Variable name	Definition	Variable name	Definition
30-day readmission	Two sequential encounters within 30 days	Diabetes	At least one of the following billing ICD-9 codes, either primary or secondary: 250.*, 648.0*
Heart failure	At least one of the following billing ICD-9 codes, either primary or secondary: 402.91, 404.01, 404.03, 404.11, 404.13, 404.91, 404.93, 428.*	Chronic kidney disease (CKD)	At least one of the following billing ICD-9 codes, either primary or secondary: 581.*, 582.*,585.*
Stroke	At least one of the following billing ICD-9 codes, either primary or secondary: 430.*, 431.*, 432.9*, 433.01, 433.11, 433.21, 433.31, 433.81, 433.91, 434.01, 434.01, 434.10, 434.90, 434.91, 435.*, 436.*	Sickle cell anemia	At least one of the following billing ICD-9 codes, either primary or secondary: 282.6*
Sickle cell crisis	At least one of the following billing ICD-9 codes, either primary or secondary: 282.62, 282.64	Frequent-flier	A patient with >= four 30-day readmissions
End-stage renal disease (ESRD)	At least one of the following billing ICD-9 codes, either primary or secondary: 285.21, 585.6	Methicillin-resistance staph aureus (MRSA)	At least one of the following billing ICD-9 codes, either primary or secondary: 041.12, 038.12
Obesity	1) At least one of the following ICD-9 codes, either primary or secondary: 278.00, 278.01; -or- 2) Body Mass Index > 30	Myocardial infarction	At least one of the following billing ICD-9 codes, either primary or secondary: 410.*
Long stayer	A patient with number of hospital days in the 75th percentile or above	Uncontrolled diabetes	1) At least one of the following ICD-9 codes, either primary or secondary: 25*.02, 25*.03, 707.1; -or- 2) HbA1c > 9%
Chemotherapy encounter	Primary or secondary billing diagnosis code V58.1*	Radiation therapy encounter	Primary or secondary billing diagnosis code V58.0
Pressure ulcer	At least one of the following ICD-9 codes, either primary or secondary: 707.0, 707.2	Rehabilitation encounter	Organization is one of the following: REH^E^SPCH, REH^W^SPCH, REH^C^SPCH, REH^E^GRPH, REH^W^GRPH, REH^C^GRPH
Planned readmit	A 30-day readmission that is a chemotherapy encounter, radiation therapy encounter, or rehabilitation encounter. These are excluded from many of our analysis.	Multiple MI	More than 1 Myocardial infarction across all encounters for a patient.
Stroke	At least one of the following billing ICD-9 codes, either primary or secondary: 430.*, 431.*, 432.9, 433.01, 433.11, 433.21, 433.31, 433.81, 433.91, 434.00, 434.01, 434.10, 434.90, 434.91, 435,*, 436.*	Cancer	At least one of the following billing ICD-9 codes, either primary or secondary: 140-208, 209.0, 209.1, 209.2, 209.3, 225.*, 227.3, 227.4, 227.9, 228.02, 228.1, 230.*, 231.*, 232.*, 233.*, 234.*, 236.0, 238.4, 238.6, 238.7, 239.6, 239.7, 259.2, 259.8, 273.2, 273.3, 285.22, 288.3, 289.83, 289.89, 511.81, 789.51, 795.06, 795.16, V58.0, V58.1*, V10.*
Chronic obstructive pulmonary disease	At least one of the following billing ICD-9 codes, either primary or secondary: 491.20, 491.21, 491.22, 492.8, 493.20, 493.21, 493.22, 494.0, 494.1, 495.*, 496.*	Pulmonary hypertension	At least one of the following billing ICD-9 codes, either primary or secondary: 416.0, 416.1, 416.8, 416.9
Metastasis	At least one of the following billing ICD-9 codes, either primary or secondary: 196.*, 197.*, 198.*	Pressure ulcer	At least one of the following billing ICD-9 codes, either primary or secondary: 707.0, 707.2
Heart failure from BNP	Standard interpretation of the B-type natriuretic peptide (BNP) laboratory test: < 100: no heart failure; 100—300: suggest heart failure is present; 300—600: mild heart failure; 600—900: moderate heart failure; > 900: severe heart failure	Multiple readmits	More than one 30-day readmission for a patient in the dataset.
UHC product lines	Derived for Emory CDW data from the MS-DRG codes that define them.		

Table A.2 Cancer Patient Characteristics

Variable	Level	Readmitted n (%) or mean±std	Non-readmitted n (%) or mean±std	Simple test* p-value
Amputation Indicator	Yes	163 (1.8)	527 (1.3)	0.0004
	No	9137 (98.2)	40564 (98.7)	
Bone Marrow Transplant Indicator	Yes	175 (1.9)	824 (2.0)	0.4401
	No	9125 (98.1)	40267 (98.0)	
Encounter 180 Days Earlier	Yes	4834 (52.0)	13152 (32.0)	<.0001
	No	4466 (48.0)	27939 (68.0)	
Encounter 90 Days Earlier	Yes	4289 (46.1)	10858 (26.4)	<.0001
	No	5011 (53.9)	30233 (73.6)	
End Stage Renal Disease Indicator	Yes	545 (5.9)	1576 (3.8)	<.0001
	No	8755 (94.1)	39515 (96.2)	
Insurance Flag	Yes	9028 (97.1)	39906 (97.1)	0.8317
	No	272 (2.9)	1185 (2.9)	
Metastasis Indicator	Yes	2183 (23.5)	9049 (22.0)	0.0024
	No	7117 (76.5)	32042 (78.0)	
Methicillin-resistant Staph Aureus Indicator	Yes	68 (0.7)	213 (0.5)	0.0128
	No	9232 (99.3)	40878 (99.5)	
Multiple MIs	Yes	186 (2.0)	323 (0.8)	<.0001
	No	9114 (98.0)	40768 (99.2)	
Multiple Readmits In the Past	Yes	1762 (18.9)	3409 (8.3)	<.0001
	No	7538 (81.1)	37682 (91.7)	
Obesity Indicator	Yes	2560 (27.5)	11146 (27.1)	0.4318
	No	6740 (72.5)	29945 (72.9)	
Sex	Male	4793 (51.5)	20753 (50.5)	0.0724
	Female	4507 (48.5)	20337 (49.5)	
Race	Asian	15 (0.2)	64 (0.2)	<.0001
	Black	3904 (42.0)	13958 (34.0)	
	Other	460 (4.9)	2729 (6.6)	
	White	4921 (52.9)	24340 (59.2)	
Pressure Ulcer Indicator	Yes	311 (3.3)	989 (2.4)	<.0001
	No	8989 (96.7)	40102 (97.6)	
Readmit Neutropenia Flag	Yes	644 (6.9)	425 (1.0)	<.0001
	No	8656 (93.1)	40666 (99.0)	
Sickle Cell Anemia Indicator	Yes	18 (0.2)	75 (0.2)	0.8230
	No	9282 (99.8)	41016 (99.8)	
Sickle Cell Crisis Indicator	Yes	10 (0.1)	54 (0.1)	0.5591
	No	9290 (99.9)	41037 (99.9)	
Uncontrolled Diabetes Indicator	Yes	347 (3.7)	1170 (2.8)	<.0001
	No	8953 (96.3)	39921 (97.2)	
Age		59.4 ± 15.8	61.6 ± 14.9	<.0001
ERAT Count		1.6 ± 0.8	1.6 ± 0.8	0.0004
Length of Stay		8.4 ± 9.5	6.9 ± 8.6	<.0001

*Two-sample t tests were performed for continuous variables and chi-square independence tests were performed for categorical variables. When less than 80% cells show frequencies >5, Fisher's exact tests were carried out instead of chi-square tests.

Table A.3 Chronic Kidney Disease Patient Characteristics

Variable	Level	Readmitted n (%) or mean±std	Non-readmitted n (%) or mean±std	Simple test* p-value
Amputation Indicator	Yes	578 (6.3)	1545 (4.7)	<.0001
	No	8652 (93.7)	31126 (95.3)	
Bone Marrow Transplant Indicator	Yes	12 (0.1)	67 (0.2)	0.1421
	No	9218 (99.9)	32604 (99.8)	
Encounter 180 Days Earlier	Yes	5645 (61.2)	13544 (41.5)	<.0001
	No	3585 (38.8)	19127 (58.5)	
Encounter 90 Days Earlier	Yes	4863 (52.7)	10705 (32.8)	<.0001
	No	4367 (47.3)	21966 (67.2)	
End Stage Renal Disease Indicator	Yes	4608 (49.9)	13055 (40.0)	<.0001
	No	4622 (50.1)	19616 (60.0)	
Insurance Flag	Yes	9047 (98.0)	31646 (96.9)	<.0001
	No	183 (2.0)	1025 (3.1)	
Metastasis Indicator	Yes	214 (2.3)	877 (2.7)	0.0513
	No	9016 (97.7)	31794 (97.3)	
Methicillin-resistant Staph Aureus Indicator	Yes	146 (1.6)	392 (1.2)	0.0040
	No	9084 (98.4)	32279 (98.8)	
Multiple MIs	Yes	575 (6.2)	1073 (3.3)	<.0001
	No	8655 (93.8)	31598 (96.7)	
Multiple Readmits In the Past	Yes	3068 (33.2)	5149 (15.8)	<.0001
	No	6162 (66.8)	27522 (84.2)	
Obesity Indicator	Yes	3285 (35.6)	12119 (37.1)	0.0082
	No	5945 (64.4)	20552 (62.9)	
Sex	Male	4757 (51.5)	17889 (54.8)	<.0001
	Female	4473 (48.5)	14781 (45.2)	
Race	Asian	14 (0.2)	84 (0.3)	<.0001
	Black	6497 (70.4)	20948 (64.1)	
	Other	270 (2.9)	1197 (3.7)	
	White	2449 (26.5)	10442 (32.0)	
Pressure Ulcer Indicator	Yes	557 (6.0)	1477 (4.5)	<.0001
	No	8673 (94.0)	31194 (95.5)	
Readmit Neutropenia Flag	Yes	61 (0.7)	58 (0.2)	<.0001
	No	9169 (99.3)	32613 (99.8)	
Sickle Cell Anemia Indicator	Yes	80 (0.9)	172 (0.5)	0.0002
	No	9150 (99.1)	32499 (99.5)	
Sickle Cell Crisis Indicator	Yes	54 (0.6)	89 (0.3)	<.0001
	No	9176 (99.4)	32582 (99.7)	
Uncontrolled Diabetes Indicator	Yes	969 (10.5)	3004 (9.2)	0.0002
	No	8261 (89.5)	29667 (90.8)	
Age		58.9 ± 16.7	61.7 ± 15.7	<.0001
ERAT Count		2.0 ± 0.9	2.0 ± 0.9	0.0065
Length of Stay		8.5 ± 10.4	7.7 ± 10.9	<.0001

*Two-sample t tests were performed for continuous variables and chi-square independence tests were performed for categorical variables. When less than 80% cells show frequencies >5, Fisher's exact tests were carried out instead of chi-square tests.

Table A.4 Chronic Obstructive Pulmonary Disease Patient Characteristics

Variable	Level	Readmitted n (%) or mean±std	Non-readmitted n (%) or mean±std	Simple test* p-value
Amputation Indicator	Yes	138 (3.7)	407 (2.3)	<.0001
	No	3567 (96.3)	17091 (97.7)	<.0001
Bone Marrow Transplant Indicator	Yes	2 (0.1)	13 (0.1)	1.0000[#]
	No	3703 (99.9)	17485 (99.9)	0.6727
Encounter 180 Days Earlier	Yes	2005 (54.1)	5484 (31.3)	<.0001
	No	1700 (45.9)	12014 (68.7)	<.0001
Encounter 90 Days Earlier	Yes	1687 (45.5)	4323 (24.7)	<.0001
	No	2018 (54.5)	13175 (75.3)	<.0001
End Stage Renal Disease Indicator	Yes	502 (13.5)	1127 (6.4)	<.0001
	No	3203 (86.5)	16371 (93.6)	<.0001
Insurance Flag	Yes	3630 (98.0)	16918 (96.7)	<.0001
	No	75 (2.0)	580 (3.3)	<.0001
Metastasis Indicator	Yes	201 (5.4)	839 (4.8)	0.1066
	No	3504 (94.6)	16659 (95.2)	0.1066
Methicillin-resistant Staph Aureus Indicator	Yes	47 (1.3)	130 (0.7)	0.0014
	No	3658 (98.7)	17368 (99.3)	0.0014
Multiple MIs	Yes	218 (5.9)	440 (2.5)	<.0001
	No	3487 (94.1)	17058 (97.5)	<.0001
Multiple Readmits In the Past	Yes	947 (25.6)	1695 (9.7)	<.0001
	No	2758 (74.4)	15803 (90.3)	<.0001
Obesity Indicator	Yes	1296 (35.0)	6031 (34.5)	0.5509
	No	2409 (65.0)	11467 (65.5)	0.5509
Sex	Male	1815 (49.0)	8804 (50.3)	0.1424
	Female	1890 (51.0)	8694 (49.7)	0.1424
Race	Asian	1 (0)	18 (0.1)	<.0001
	Black	1821 (49.1)	6538 (37.4)	<.0001
	Other	96 (2.6)	705 (4.0)	<.0001
	White	1787 (48.2)	10237 (58.5)	<.0001
Pressure Ulcer Indicator	Yes	188 (5.1)	579 (3.3)	<.0001
	No	3517 (94.9)	16919 (96.7)	<.0001
Readmit Neutropenia Flag	Yes	16 (0.4)	29 (0.2)	0.0014
	No	3689 (99.6)	17469 (99.8)	0.0014
Sickle Cell Anemia Indicator	Yes	24 (0.6)	78 (0.4)	0.1064
	No	3681 (99.4)	17420 (99.6)	0.1064
Sickle Cell Crisis Indicator	Yes	23 (0.6)	68 (0.4)	0.0496
	No	3682 (99.4)	17430 (99.6)	0.0496
Uncontrolled Diabetes Indicator	Yes	304 (8.2)	1067 (6.1)	<.0001
	No	3401 (91.8)	16431 (93.9)	<.0001
Age		65.8 ± 12.6	66.5 ± 12.0	0.0052
ERAT Count		2.3 ± 1.0	2.1 ± 1.0	<.0001
Length of Stay		8.8 ± 10.8	7.3 ± 8.6	<.0001

*Two-sample t tests were performed for continuous variables and chi-square independence tests were performed for categorical variables. When less than 80% cells show frequencies >5, Fisher's exact tests were carried out instead of chi-square tests.
[#]Fisher's exact test.

Table A.5 Diabetes Patient Characteristics

Variable	Level	Readmitted n (%) or mean±std	Non-readmitted n (%) or mean±std	Simple test* p-value
Amputation Indicator	Yes	699 (7.5)	2170 (4.7)	<.0001
	No	8593 (92.5)	43631 (95.3)	
Bone Marrow Transplant Indicator	Yes	16 (0.2)	104 (0.2)	0.3009
	No	9276 (99.8)	45697 (99.8)	
Encounter 180 Days Earlier	Yes	5005 (53.9)	14205 (31.0)	<.0001
	No	4287 (46.1)	31596 (69.0)	
Encounter 90 Days Earlier	Yes	4301 (46.3)	11135 (24.3)	<.0001
	No	4991 (53.7)	34666 (75.7)	
End Stage Renal Disease Indicator	Yes	2044 (22.0)	6071 (13.3)	<.0001
	No	7248 (78.0)	39730 (86.7)	
Insurance Flag	Yes	8979 (96.6)	43697 (95.4)	<.0001
	No	313 (3.4)	2104 (4.6)	
Metastasis Indicator	Yes	423 (4.6)	1686 (3.7)	<.0001
	No	8869 (95.4)	44115 (96.3)	
Methicillin-resistant Staph Aureus Indicator	Yes	133 (1.4)	412 (0.9)	<.0001
	No	9159 (98.6)	45389 (99.1)	
Multiple MIs	Yes	580 (6.2)	1316 (2.9)	<.0001
	No	8712 (93.8)	44485 (97.1)	
Multiple Readmits In the Past	Yes	2380 (25.6)	4280 (9.3)	<.0001
	No	6912 (74.4)	41521 (90.7)	
Obesity Indicator	Yes	4173 (44.9)	21393 (46.7)	0.0015
	No	5119 (55.1)	24408 (53.3)	
Sex	Male	4330 (46.6)	22147 (48.4)	0.0020
	Female	4962 (53.4)	23652 (51.6)	
Race	Asian	17 (0.2)	119 (0.3)	<.0001
	Black	5373 (57.8)	23753 (51.9)	
	Other	362 (3.9)	2462 (5.4)	
	White	3540 (38.1)	19467 (42.5)	
Pressure Ulcer Indicator	Yes	607 (6.5)	1720 (3.8)	<.0001
	No	8685 (93.5)	44081 (96.2)	
Readmit Neutropenia Flag	Yes	91 (1.0)	93 (0.2)	<.0001
	No	9201 (99.0)	45708 (99.8)	
Sickle Cell Anemia Indicator	Yes	18 (0.2)	50 (0.1)	0.0343
	No	9274 (99.8)	45751 (99.9)	
Sickle Cell Crisis Indicator	Yes	15 (0.2)	27 (0.1)	0.0011
	No	9277 (99.8)	45774 (99.9)	
Uncontrolled Diabetes Indicator	Yes	1916 (20.6)	8367 (18.3)	<.0001
	No	7376 (79.4)	37434 (81.7)	
Age		61.4 ± 14.4	62.5 ± 13.8	<.0001
ERAT Count		2.1 ± 0.9	1.9 ± 0.9	<.0001
Length of Stay		8.3 ± 10.3	6.5 ± 9.5	<.0001

*Two-sample t tests were performed for continuous variables and chi-square independence tests were performed for categorical variables. When less than 80% cells show frequencies >5, Fisher's exact tests were carried out instead of chi-square tests.

Table A.6 Heart Failure Patient Characteristics

Variable	Level	Readmitted n (%) or mean±std	Non-readmitted n (%) or mean±std	Simple test* p-value
Amputation Indicator	Yes	259 (5.0)	644 (3.9)	0.0010
	No	4938 (95.0)	15709 (96.1)	
Bone Marrow Transplant Indicator	Yes	3 (0.1)	16 (0.1)	0.5918[#]
	No	5194 (99.9)	16337 (99.9)	
Encounter 180 Days Earlier	Yes	4283 (82.4)	11340 (69.3)	<.0001
	No	914 (17.6)	5013 (30.7)	
Encounter 90 Days Earlier	Yes	3654 (70.3)	8954 (54.8)	<.0001
	No	1543 (29.7)	7399 (45.2)	
End Stage Renal Disease Indicator	Yes	1191 (22.9)	2651 (16.2)	<.0001
	No	4006 (77.1)	13702 (83.8)	
Insurance Flag	Yes	5060 (97.4)	15793 (96.6)	0.0051
	No	137 (2.6)	560 (3.4)	
Metastasis Indicator	Yes	122 (2.3)	359 (2.2)	0.5176
	No	5075 (97.7)	15994 (97.8)	
Methicillin-resistant Staph Aureus Indicator	Yes	69 (1.3)	164 (1.0)	0.0486
	No	5128 (98.7)	16189 (99.0)	
Multiple MIs	Yes	434 (8.4)	1083 (6.6)	<.0001
	No	4763 (91.6)	15270 (93.4)	
Multiple Readmits In the Past	Yes	2053 (39.5)	3790 (23.2)	<.0001
	No	3144 (60.5)	12563 (76.8)	
Obesity Indicator	Yes	2125 (40.9)	6794 (41.5)	0.4023
	No	3072 (59.1)	9559 (58.5)	
Sex	Male	2564 (49.3)	8180 (50.0)	0.3894
	Female	2633 (50.7)	8173 (50.0)	
Race	Asian	7 (0.1)	34 (0.2)	<.0001
	Black	3418 (65.8)	9730 (59.5)	
	Other	122 (2.3)	426 (2.6)	
	White	1650 (31.7)	6163 (37.7)	
Pressure Ulcer Indicator	Yes	325 (6.3)	816 (5.0)	0.0004
	No	4872 (93.7)	15537 (95.0)	
Readmit Neutropenia Flag	Yes	36 (0.7)	22 (0.1)	<.0001
	No	5161 (99.3)	16331 (99.9)	
Sickle Cell Anemia Indicator	Yes	46 (0.9)	79 (0.5)	0.0009
	No	5151 (99.1)	16274 (99.5)	
Sickle Cell Crisis Indicator	Yes	36 (0.7)	49 (0.3)	<.0001
	No	5161 (99.3)	16304 (99.7)	
Uncontrolled Diabetes Indicator	Yes	534 (10.3)	1489 (9.1)	0.0118
	No	4663 (89.7)	14864 (90.9)	
Age		61.8 ± 16.2	64.1 ± 15.2	<.0001
ERAT Count		1.8 ± 1.0	1.7 ± 1.1	<.0001
Length of Stay		8.3 ± 10.6	7.4 ± 10.7	<.0001

*Two-sample t tests were performed for continuous variables and chi-square independence tests were performed for categorical variables. When less than 80% cells show frequencies >5, Fisher's exact tests were carried out instead of chi-square tests.
[#]Fisher's exact test.

Table A.7 Acute Myocardial Infarction Patient Characteristics

Variable	Level	Readmitted n (%) or mean±std	Non-readmitted n (%) or mean±std	Simple test* p-value
Amputation Indicator	Yes	195 (6.3)	528 (2.7)	<.0001
	No	2887 (93.7)	18793 (97.3)	
Bone Marrow Transplant Indicator	Yes	4 (0.1)	13 (0.1)	0.2783#
	No	3078 (99.9)	19308 (99.9)	
Encounter 180 Days Earlier	Yes	1386 (45.0)	4581 (23.7)	<.0001
	No	1696 (55.0)	14740 (76.3)	
Encounter 90 Days Earlier	Yes	1180 (38.3)	3622 (18.7)	<.0001
	No	1902 (61.7)	15699 (81.3)	
End Stage Renal Disease Indicator	Yes	504 (16.4)	1532 (7.9)	<.0001
	No	2578 (83.6)	17789 (92.1)	
Insurance Flag	Yes	2943 (95.5)	17744 (91.8)	<.0001
	No	139 (4.5)	1577 (8.2)	
Metastasis Indicator	Yes	113 (3.7)	434 (2.2)	<.0001
	No	2969 (96.3)	18887 (97.8)	
Methicillin-resistant Staph Aureus Indicator	Yes	35 (1.1)	94 (0.5)	<.0001
	No	3047 (98.9)	19227 (99.5)	
Multiple MIs	Yes	982 (31.9)	1979 (10.2)	<.0001
	No	2100 (68.1)	17342 (89.8)	
Multiple Readmits In the Past	Yes	560 (18.2)	1177 (6.1)	<.0001
	No	2522 (81.8)	18144 (93.9)	
Obesity Indicator	Yes	1127 (36.6)	6870 (35.6)	0.2771
	No	1955 (63.4)	12451 (64.4)	
Sex	Male	1725 (56.0)	11828 (61.2)	<.0001
	Female	1357 (44.0)	7493 (38.8)	
Race	Asian	5 (0.2)	42 (0.2)	<.0001
	Black	1511 (49.0)	7267 (37.6)	
	Other	116 (3.8)	1153 (6.0)	
	White	1450 (47.0)	10859 (56.2)	
Pressure Ulcer Indicator	Yes	158 (5.1)	456 (2.4)	<.0001
	No	2924 (94.9)	18865 (97.6)	
Readmit Neutropenia Flag	Yes	25 (0.8)	15 (0.1)	<.0001
	No	3057 (99.2)	19306 (99.9)	
Sickle Cell Anemia Indicator	Yes	7 (0.2)	23 (0.1)	0.1276
	No	3075 (99.8)	19298 (99.9)	
Sickle Cell Crisis Indicator	Yes	4 (0.1)	11 (0.1)	0.1407#
	No	3078 (99.9)	19310 (99.9)	
Uncontrolled Diabetes Indicator	Yes	352 (11.4)	1501 (7.8)	<.0001
	No	2730 (88.6)	17820 (92.2)	
Age		64.4 ± 13.1	64.4 ± 12.9	0.9484
ERAT Count		2.3 ± 1.0	2.0 ± 1.0	<.0001
Length of Stay		8.4 ± 9.9	6.3 ± 8.3	<.0001

*Two-sample t tests were performed for continuous variables and chi-square independence tests were performed for categorical variables. When less than 80% cells show frequencies >5, Fisher's exact tests were carried out instead of chi-square tests.
#Fisher's exact test.

Table A.8 Pulmonary Hypertension Patient Characteristics

Variable	Level	Readmitted n (%) or mean±std	Non-readmitted n (%) or mean±std	Simple test* p-value
Amputation Indicator	Yes	36 (2.9)	93 (2.0)	0.0539
	No	1222 (97.1)	4622 (98.0)	
Bone Marrow Transplant Indicator	Yes	1 (0.1)	2 (0)	0.5082[#]
	No	1257 (99.9)	4713 (100.0)	
Encounter 180 Days Earlier	Yes	764 (60.7)	1866 (39.6)	<.0001
	No	494 (39.3)	2849 (60.4)	
Encounter 90 Days Earlier	Yes	644 (51.2)	1469 (31.2)	<.0001
	No	614 (48.8)	3246 (68.8)	
End Stage Renal Disease Indicator	Yes	273 (21.7)	639 (13.6)	<.0001
	No	985 (78.3)	4076 (86.4)	
Insurance Flag	Yes	1226 (97.5)	4561 (96.7)	0.1900
	No	32 (2.5)	154 (3.3)	
Metastasis Indicator	Yes	20 (1.6)	68 (1.4)	0.6994
	No	1238 (98.4)	4647 (98.6)	
Methicillin-resistant Staph Aureus Indicator	Yes	17 (1.4)	28 (0.6)	0.0058
	No	1241 (98.6)	4687 (99.4)	
Multiple MIs	Yes	55 (4.4)	115 (2.4)	0.0002
	No	1203 (95.6)	4600 (97.6)	
Multiple Readmits In the Past	Yes	363 (28.9)	674 (14.3)	<.0001
	No	895 (71.1)	4041 (85.7)	
Obesity Indicator	Yes	515 (40.9)	2017 (42.8)	0.2406
	No	743 (59.1)	2698 (57.2)	
Sex	Male	445 (35.4)	1702 (36.1)	0.6345
	Female	813 (64.6)	3013 (63.9)	
Race	Asian	3 (0.2)	10 (0.2)	<.0001
	Black	774 (61.5)	2567 (54.4)	
	Other	55 (4.4)	213 (4.5)	
	White	426 (33.9)	1925 (40.8)	
Pressure Ulcer Indicator	Yes	81 (6.4)	157 (3.3)	<.0001
	No	1177 (93.6)	4558 (96.7)	
Readmit Neutropenia Flag	Yes	6 (0.5)	7 (0.1)	0.0263
	No	1252 (99.5)	4708 (99.9)	
Sickle Cell Anemia Indicator	Yes	40 (3.2)	115 (2.4)	0.1421
	No	1218 (96.8)	4600 (97.6)	
Sickle Cell Crisis Indicator	Yes	35 (2.8)	99 (2.1)	0.1464
	No	1223 (97.2)	4616 (97.9)	
Uncontrolled Diabetes Indicator	Yes	121 (9.6)	345 (7.3)	0.0069
	No	1137 (90.4)	4370 (92.7)	
Age		59.3 ± 17.4	61.8 ± 16.4	<.0001
ERAT Count		2.5 ± 1.0	2.4 ± 1.1	0.0013
Length of Stay		10.9 ± 14.3	9.0 ± 11.5	<.0001

*Two-sample t tests were performed for continuous variables and chi-square independence tests were performed for categorical variables. When less than 80% cells show frequencies >5, Fisher's exact tests were carried out instead of chi-square tests.
[#]Fisher's exact test.

Table A.9 Sickle Cell Anemia Patient Characteristics

Variable	Level	Readmitted n (%) or mean±std	Non-readmitted n (%) or mean±std	Simple test* p-value
Amputation Indicator	Yes	1 (0.1)	2 (0.1)	1.0000[#]
	No	744 (99.9)	1439 (99.9)	
Bone Marrow Transplant Indicator	Yes	1 (0.1)	0 (0)	0.3408[#]
	No	744 (99.9)	1441 (100.0)	
Encounter 180 Days Earlier	Yes	659 (88.5)	923 (64.1)	<.0001
	No	86 (11.5)	518 (35.9)	
Encounter 90 Days Earlier	Yes	616 (82.7)	729 (50.6)	
	No	129 (17.3)	712 (49.4)	
End Stage Renal Disease Indicator	Yes	52 (7.0)	93 (6.5)	0.6395
	No	693 (93.0)	1348 (93.5)	
Insurance Flag	Yes	730 (98.0)	1381 (95.8)	0.0088
	No	15 (2.0)	60 (4.2)	
Metastasis Indicator	Yes	1 (0.1)	5 (0.3)	0.6706[#]
	No	744 (99.9)	1436 (99.7)	
Methicillin-resistant Staph Aureus Indicator	Yes	12 (1.6)	9 (0.6)	0.0251
	No	733 (98.4)	1432 (99.4)	
Multiple MIs	Yes	0 (0)	0 (0)	
	No	745 (100.0)	1441 (100.0)	
Multiple Readmits In the Past	Yes	526 (70.6)	513 (35.6)	<.0001
	No	219 (29.4)	928 (64.4)	
Obesity Indicator	Yes	99 (13.3)	223 (15.5)	0.1715
	No	646 (86.7)	1218 (84.5)	
Sex	Male	315 (42.3)	523 (36.3)	0.0064
	Female	430 (57.7)	918 (63.7)	
Race	Asian	1 (0.1)	0 (0)	0.0316[#]
	Black	743 (99.7)	1426 (99.0)	
	Other	0 (0)	8 (0.6)	
	White	1 (0.1)	7 (0.5)	
Pressure Ulcer Indicator	Yes	1 (0.1)	0 (0)	0.3408[#]
	No	744 (99.9)	1441 (100.0)	
Readmit Neutropenia Flag	Yes	2 (0.3)	0 (0)	0.1160[#]
	No	743 (99.7)	1441 (100.0)	
Sickle Cell Anemia Indicator	Yes	745 (100.0)	1441 (100.0)	
	No	0 (0)	0 (0)	
Sickle Cell Crisis Indicator	Yes	660 (88.6)	1145 (79.5)	<.0001
	No	85 (11.4)	296 (20.5)	
Uncontrolled Diabetes Indicator	Yes	40 (5.4)	64 (4.4)	0.3341
	No	705 (94.6)	1377 (95.6)	
Age		32.0 ± 10.9	35.6 ± 12.7	<.0001
ERAT Count		0.3 ± 0.5	0.4 ± 0.6	<.0001
Length of Stay		7.1 ± 6.3	6.7 ± 7.2	0.1180

*Two-sample t tests were performed for continuous variables and chi-square independence tests were performed for categorical variables. When less than 80% cells show frequencies >5, Fisher's exact tests were carried out instead of chi-square tests.
[#]Fisher's exact test.

Table A.10 Stroke Patient Characteristics

Variable	Level	Readmitted n (%) or mean±std	Non-readmitted n (%) or mean±std	Simple test* p-value
Amputation Indicator	Yes	24 (2.9)	90 (1.5)	0.004
	No	818 (97.2)	5926 (98.5)	
Bone Marrow Transplant Indicator	Yes	0 (0)	0 (0)	
	No	842 (100.0)	6016 (100.0)	
Encounter 180 Days Earlier	Yes	212 (25.2)	925 (15.4)	<.0001
	No	630 (74.8)	5091 (84.6)	
Encounter 90 Days Earlier	Yes	174 (20.7)	741 (12.3)	<.0001
	No	668 (79.3)	5275 (87.7)	
End Stage Renal Disease Indicator	Yes	49 (5.8)	297 (4.9)	0.2731
	No	793 (94.2)	5719 (95.1)	
Insurance Flag	Yes	777 (92.3)	5428 (90.2)	0.0572
	No	65 (7.7)	588 (9.8)	
Metastasis Indicator	Yes	25 (3.0)	184 (3.1)	0.8876
	No	817 (97.0)	5832 (96.9)	
Methicillin-resistant Staph Aureus Indicator	Yes	13 (1.5)	26 (0.4)	<.0001
	No	829 (98.5)	5990 (99.6)	
Multiple MIs	Yes	28 (3.3)	62 (1.0)	<.0001
	No	814 (96.7)	5954 (99.0)	
Multiple Readmits In the Past	Yes	55 (6.5)	219 (3.6)	<.0001
	No	787 (93.5)	5797 (96.4)	
Obesity Indicator	Yes	298 (35.4)	1947 (32.4)	0.0795
	No	544 (64.6)	4069 (67.6)	
Sex	Male	369 (43.8)	2726 (45.3)	0.4163
	Female	473 (56.2)	3290 (54.7)	
Race	Asian	3 (0.4)	4 (0.1)	<.0001
	Black	422 (50.1)	2854 (47.4)	
	Other	50 (5.9)	708 (11.8)	
	White	367 (43.6)	2450 (40.7)	
Pressure Ulcer Indicator	Yes	51 (6.1)	159 (2.6)	<.0001
	No	791 (93.9)	5857 (97.4)	
Readmit Neutropenia Flag	Yes	0 (0)	5 (0.1)	1.0000[#]
	No	842 (100.0)	6011 (99.9)	
Sickle Cell Anemia Indicator	Yes	7 (0.8)	11 (0.2)	0.0006
	No	835 (99.2)	6005 (99.8)	
Sickle Cell Crisis Indicator	Yes	5 (0.6)	6 (0.1)	0.0008
	No	837 (99.4)	6010 (99.9)	
Uncontrolled Diabetes Indicator	Yes	57 (6.8)	351 (5.8)	0.2826
	No	785 (93.2)	5665 (94.2)	
Age		60.4 ± 15.5	61.6 ± 15.6	0.0397
ERAT Count		1.9 ± 0.9	1.8 ± 0.9	0.0414
Length of Stay		13.3 ± 13.4	9.1 ± 14.1	<.0001

*Two-sample t tests were performed for continuous variables and chi-square independence tests were performed for categorical variables. When less than 80% cells show frequencies >5, Fisher's exact tests were carried out instead of chi-square tests.
[#]Fisher's exact test.

Table A.11 Transplant Patient Characteristics

Variable	Level	Readmitted n (%) or mean±std	Non-readmitted n (%) or mean±std	Simple test* p-value
Amputation Indicator	Yes	41 (3.5)	116 (2.9)	0.3073
	No	1130 (96.5)	3860 (97.1)	
Bone Marrow Transplant Indicator	Yes	4 (0.3)	25 (0.6)	0.3726[#]
	No	1167 (99.7)	3951 (99.4)	
Encounter 180 Days Earlier	Yes	825 (70.5)	2067 (52.0)	<.0001
	No	346 (29.5)	1909 (48.0)	
Encounter 90 Days Earlier	Yes	723 (61.7)	1659 (41.7)	<.0001
	No	448 (38.3)	2317 (58.3)	
End Stage Renal Disease Indicator	Yes	146 (12.5)	390 (9.8)	0.0088
	No	1025 (87.5)	3586 (90.2)	
Insurance Flag	Yes	1150 (98.2)	3885 (97.7)	0.3071
	No	21 (1.8)	91 (2.3)	
Metastasis Indicator	Yes	15 (1.3)	85 (2.1)	0.0619
	No	1156 (98.7)	3891 (97.9)	
Methicillin-resistant Staph Aureus Indicator	Yes	12 (1.0)	30 (0.8)	0.3663
	No	1159 (99.0)	3946 (99.2)	
Multiple MIs	Yes	27 (2.3)	43 (1.1)	0.0015
	No	1144 (97.7)	3933 (98.9)	
Multiple Readmits In the Past	Yes	421 (36.0)	881 (22.2)	<.0001
	No	750 (64.0)	3095 (77.8)	
Obesity Indicator	Yes	289 (24.7)	1035 (26.0)	0.3524
	No	882 (75.3)	2941 (74.0)	
Sex	Male	694 (59.3)	2227 (56.0)	0.0482
	Female	477 (40.7)	1749 (44.0)	
Race	Asian	2 (0.2)	12 (0.3)	0.0333
	Black	368 (31.4)	1220 (30.7)	
	Other	49 (4.2)	255 (6.4)	
	White	752 (64.2)	2489 (62.6)	
Pressure Ulcer Indicator	Yes	37 (3.2)	69 (1.7)	0.0026
	No	1134 (96.8)	3907 (98.3)	
Readmit Neutropenia Flag	Yes	40 (3.4)	39 (1.0)	<.0001
	No	1131 (96.6)	3937 (99.0)	
Sickle Cell Anemia Indicator	Yes	6 (0.5)	14 (0.4)	0.4385
	No	1165 (99.5)	3962 (99.6)	
Sickle Cell Crisis Indicator	Yes	2 (0.2)	5 (0.1)	0.4271[#]
	No	1169 (99.8)	3971 (99.9)	
Uncontrolled Diabetes Indicator	Yes	53 (4.5)	178 (4.5)	0.9430
	No	1118 (95.5)	3798 (95.5)	
Age		53.7 ± 14.4	54.6 ± 15.3	0.0508
ERAT Count		2.1 ± 0.8	2.1 ± 0.9	0.9079
Length of Stay		6.7 ± 7.0	6.0± 7.2	0.0038

*Two-sample t tests were performed for continuous variables and chi-square independence tests were performed for categorical variables. When less than 80% cells show frequencies >5, Fisher's exact tests were carried out instead of chi-square tests.
[#]Fisher's exact test.

Table A.12 Ranking of Readmission Risk by Random Forest (10 trees) Model, part I

Subpopulation	Ranking	# of encounter	# of Readmission	Readmission rate	Improvement over baseline
Cancer	>= 90%	1012	461	46%	156%
	75% - 90%	2275	546	24%	33%
	50% - 75%	1977	346	18%	0%
	25% - 50%	2566	318	12%	33%
	10% - 25%	2276	187	8%	56%
	<= 10%	0	0	NA	NA
	Total	10106	1858	18%	
Chronic Kidney Disease	>= 90%	1127	499	44%	100%
	75% - 90%	2273	618	27%	23%
	50% - 75%	1707	358	21%	5%
	25% - 50%	1871	277	15%	32%
	10% - 25%	1427	134	9%	0.59
	<= 10%	0	0	NA	NA
	Total	8405	1886	22%	
Chronic Obstructive Pulmonary Disease	>= 90%	735	275	37%	118%
	75% - 90%	593	139	23%	35%
	50% - 75%	841	129	15%	12%
	25% - 50%	1080	120	11%	35%
	10% - 25%	1030	84	8%	53%
	<= 10%	0	0	NA	NA
	Total	4279	747	17%	
Diabetes	>= 90%	1801	674	37%	118%
	75% - 90%	1443	283	20%	18%
	50% - 75%	4994	689	14%	18%
	25% - 50%	2801	237	8%	53%
	10% - 25%	0	0	NA	NA
	<= 10%	0	0	NA	NA
	Total	11039	1883	17%	
Heart Failure	>= 90%	685	288	42%	75%
	75% - 90%	567	145	26%	8%
	50% - 75%	1683	392	23%	4%
	25% - 50%	911	138	15%	38%
	10% - 25%	514	66	13%	46%
	<= 10%	0	0	NA	NA
	Total	4360	1029	24%	

Table A.13 Ranking of Readmission Risk by Random Forest (10 trees) Model, part II

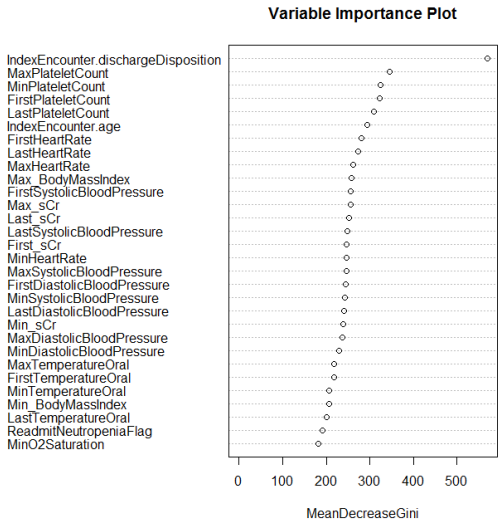
Subpopulation	Ranking	# of encounter	# of Readmission	Readmission rate	Improvement over baseline
Acute Myocardial Infarction	>= 90%	557	216	39%	179%
	75% - 90%	1210	216	18%	29%
	50% - 75%	1144	121	11%	21%
	25% - 50%	1616	77	5%	64%
	10% - 25%	0	0	NA	NA
	<= 10%	0	0	NA	NA
	Total	4527	630	14%	
Pulmonary Hypertension	>= 90%	135	51	38%	81%
	75% - 90%	305	75	25%	19%
	50% - 75%	227	59	26%	24%
	25% - 50%	282	40	14%	33%
	10% - 25%	213	19	9%	57%
	<= 10%	0	0	NA	NA
	Total	1162	244	21%	
Sickle Cell Anemia	>= 90%	53	33	62%	94%
	75% - 90%	99	43	43%	34%
	50% - 75%	102	31	30%	6%
	25% - 50%	72	12	17%	47%
	10% - 25%	57	11	19%	41%
	<= 10%	34	5	15%	53%
	Total	417	135	32%	
Stroke	>= 90%	142	65	46%	283%
	75% - 90%	297	43	14%	17%
	50% - 75%	386	31	8%	33%
	25% - 50%	532	28	5%	58%
	10% - 25%	0	0	NA	NA
	<= 10%	0	0	NA	NA
	Total	1357	167	12%	
Transplant	>= 90%	152	58	38%	65%
	75% - 90%	114	28	25%	9%
	50% - 75%	412	85	21%	9%
	25% - 50%	193	36	19%	17%
	10% - 25%	105	19	18%	22%
	<= 10%	0	0	NA	NA
	Total	976	226	23%	

Table A.14 Ranking of Readmission Risk by Random Forest (50 trees) Model, part I

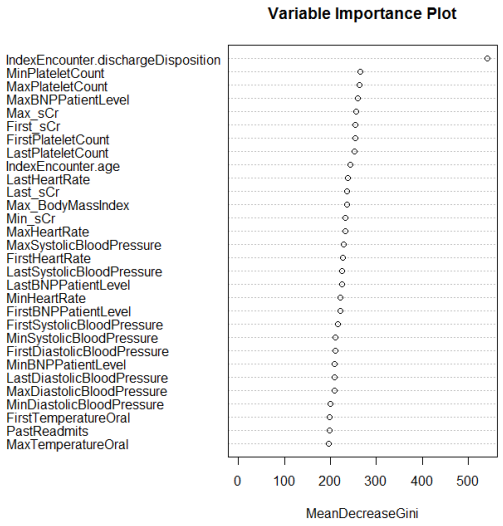
Subpopulation	Ranking	# of encounter	# of Readmission	Readmission rate	Improvement over baseline
Cancer	>= 90%	1077	550	51%	183%
	75% - 90%	1619	428	26%	44%
	50% - 75%	2894	489	17%	6%
	25% - 50%	2002	221	11%	39%
	10% - 25%	1866	139	7%	61%
	<= 10%	648	31	5%	72%
	Total	10106	1858	18%	
Chronic Kidney Disease	>= 90%	891	448	50%	127%
	75% - 90%	1383	481	35%	59%
	50% - 75%	1946	477	25%	14%
	25% - 50%	2358	339	14%	36%
	10% - 25%	1225	114	9%	0.59
	<= 10%	602	27	4%	82%
	Total	8405	1886	22%	
Chronic Obstructive Pulmonary Disease	>= 90%	496	234	47%	176%
	75% - 90%	714	166	23%	35%
	50% - 75%	1081	177	16%	6%
	25% - 50%	1135	109	10%	41%
	10% - 25%	300	26	9%	47%
	<= 10%	553	35	6%	65%
	Total	4279	747	17%	
Diabetes	>= 90%	1140	540	47%	176%
	75% - 90%	1724	456	26%	53%
	50% - 75%	2689	426	16%	6%
	25% - 50%	3087	317	10%	41%
	10% - 25%	1483	107	7%	59%
	<= 10%	916	37	4%	76%
	Total	11039	1883	17%	
Heart Failure	>= 90%	470	245	52%	117%
	75% - 90%	754	244	32%	33%
	50% - 75%	1064	259	24%	0%
	25% - 50%	990	170	17%	29%
	10% - 25%	648	77	12%	50%
	<= 10%	434	34	8%	67%
	Total	4360	1029	24%	

Table A.15 Ranking of Readmission Risk by Random Forest (50 trees) Model, part II

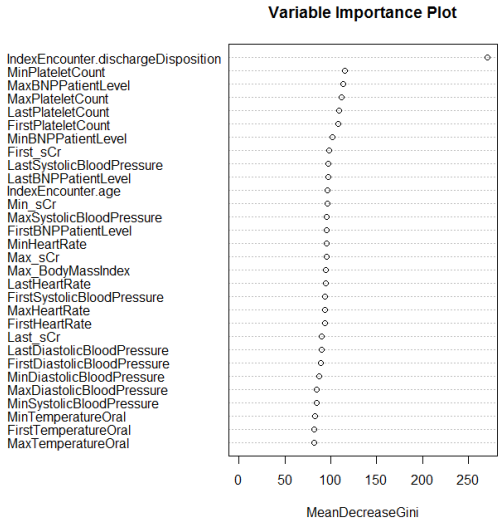
Subpopulation	Ranking	# of encounter	# of Readmission	Readmission rate	Improvement over baseline
Acute Myocardial Infarction	>= 90%	521	240	46%	229%
	75% - 90%	655	157	24%	71%
	50% - 75%	1101	134	12%	14%
	25% - 50%	1425	87	6%	57%
	10% - 25%	432	10	2%	86%
	<= 10%	393	2	1%	93%
	Total	4527	630	14%	
Pulmonary Hypertension	>= 90%	118	50	42%	100%
	75% - 90%	173	65	38%	81%
	50% - 75%	314	65	21%	0%
	25% - 50%	332	47	14%	33%
	10% - 25%	117	13	11%	48%
	<= 10%	108	4	4%	81%
	Total	1162	244	21%	
Sickle Cell Anemia	>= 90%	44	32	73%	128%
	75% - 90%	66	31	47%	47%
	50% - 75%	101	36	36%	12%
	25% - 50%	112	19	17%	47%
	10% - 25%	58	11	19%	41%
	<= 10%	36	6	17%	47%
	Total	417	135	32%	
Stroke	>= 90%	147	73	50%	317%
	75% - 90%	222	34	15%	25%
	50% - 75%	436	37	8%	33%
	25% - 50%	302	17	6%	50%
	10% - 25%	144	5	3%	75%
	<= 10%	106	1	1%	92%
	Total	1357	167	12%	
Transplant	>= 90%	109	47	43%	87%
	75% - 90%	176	49	28%	22%
	50% - 75%	208	54	26%	13%
	25% - 50%	287	49	17%	26%
	10% - 25%	100	10	10%	57%
	<= 10%	96	17	18%	22%
	Total	976	226	23%	



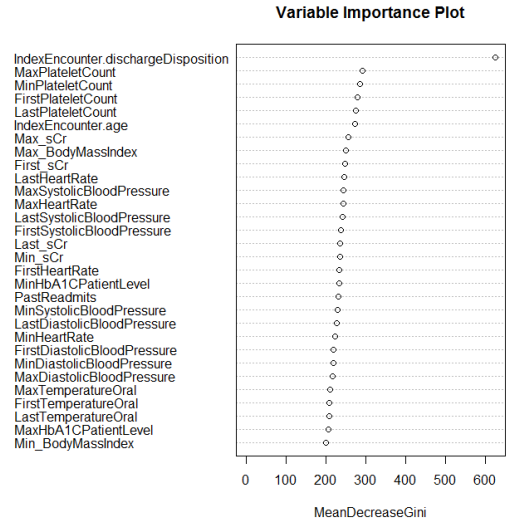
A. Cancer patients



B. Choronic Kidney Disease patients

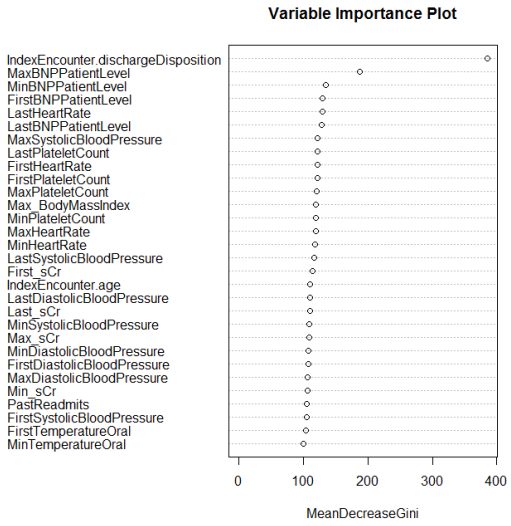


C. Chronic Obstructive Pulmonary Disease patients

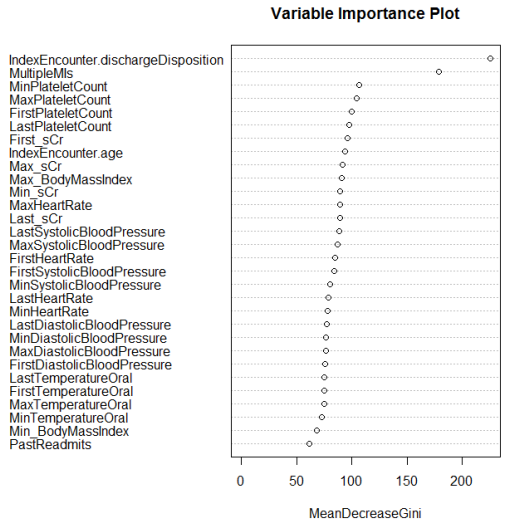


D. Diabetes patients

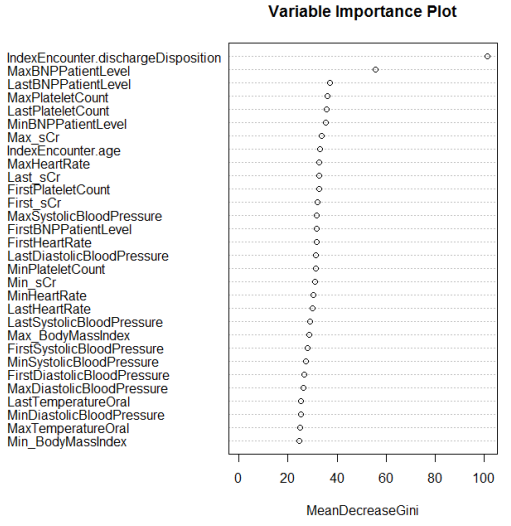
Figure A.1 List of important variables in the RF models for patient subpopulations, part I.



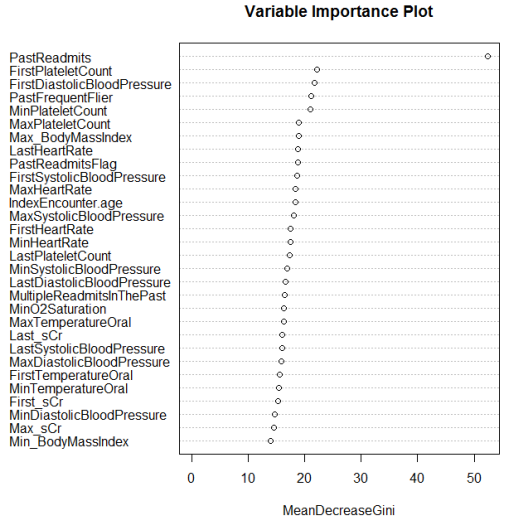
A. Heart Failure patients



B. Acute Myocardial Infarction patients

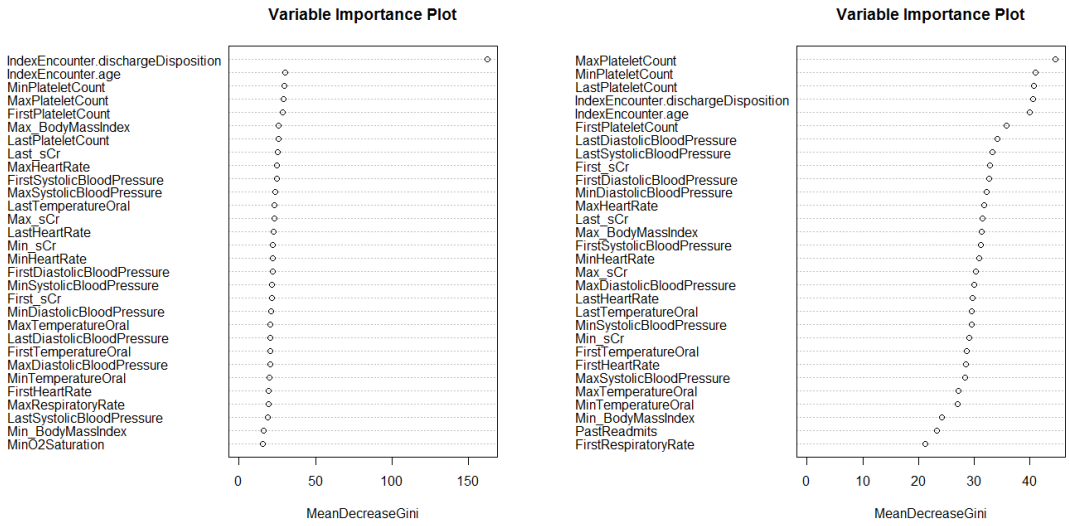


C. Pulmonary Hypertension patients



D. Sickle Cell Anemia patients

Figure A.2 List of important variables in the RF models for patient subpopulations, part II.



A. Stroke patients B. Transplant patients

Figure A.3 List of important variables in the RF models for patient subpopulations, part III.