# Predicting All-Cause Hospital Readmissions from Medical Claims Data of Hospitalised Patients

Avinash Kadimisetty*, Arun Rajagopalan† and Vijendra SK‡

Evive Software Analytics Pvt. Ltd., Bengaluru, Karnataka, India

Email: *avinash.k@goevive.com, †arun.rajagopalan@goevive.com, ‡vijendra@goevive.com

*Abstract*—Reducing preventable hospital readmissions is a national priority for payers, providers, and policymakers seeking to improve health care and lower costs. The rate of readmission is being used as a benchmark to determine the quality of healthcare provided by the hospitals. In this project, we have used machine learning techniques like Logistic Regression, Random Forest and Support Vector Machines to analyze the health claims data and identify demographic and medical factors that play a crucial role in predicting all-cause readmissions. As the health claims data is high dimensional, we have used Principal Component Analysis as a dimension reduction technique and used the results for building regression models. We compared and evaluated these models based on the Area Under Curve (AUC) metric. Random Forest model gave the highest performance followed by Logistic Regression and Support Vector Machine models. These models can be used to identify the crucial factors causing readmissions and help identify patients to focus on to reduce the chances of readmission, ultimately bringing down the cost and increasing the quality of healthcare provided to the patients.

*Index Terms*—Hospital Readmission, Comorbidity, Risk, Classification, Random Forest, Support Vector Machine.

## I. INTRODUCTION

More than a trillion dollars are being spent annually in the healthcare industry partly due to the latest technology not being used to its fullest in healthcare. Machine Learning techniques can have a huge impact in reducing the healthcare costs that are expected to increase in the coming years. Patients' readmissions to the hospitals are one of the reasons for the increasing costs in healthcare. Readmissions often occur due to poor treatment provided to the patients, more specifically, they are often caused by premature discharges or communication breakdown between patients and the healthcare team while the patient is being discharged.

Unplanned Hospital Readmission is defined as an unexpected readmission to the same hospital within 28 or 30 days of being discharged. However, the literature has widely used 30 days within the context of measurement of Hospital Readmissions. Unplanned Hospital Readmission rate is considered as a performance indicator to measure a hospital's quality of care.

Unplanned readmissions cause a disruption to the normality of the patients lives and result in significant financial burden on the healthcare system. In the USA alone, it has been estimated that 20% (7.8 million) of the hospital discharged patients were readmitted. These readmissions result in higher costs to taxpayers, costing as much as

$45 billion annually. Medicare, along with other healthcare payers, are concerned with the cost of unnecessary readmissions as Medicare alone spends roughly $15 billion annually on repeat hospitalizations. Almost 76% of the repeat admissions can be avoided by improving care before and after the patient is discharged. By decreasing these preventable repeat hospitalizations, overall productivity of the hospitals can improve considerably. In this study, we aimed at building predictive models to identify patients at high risk for readmissions. Preventive approaches can then be developed and applied to target the identified high-risk patients.

## II. LITERATURE SURVEY

As per the Affordable Care Act of 2010, hospitals reimburse a partial amount to the patients readmitted to the hospital within 30 days from discharge [8]. As these readmissions are costly and considered as an indication of poor quality, many studies have been performed to identify various factors that play crucial roles in predicting possible readmissions thereby alleviating revenue losses. As the hospital readmissions are driven by the nature of the population, a few studies involved deeper analysis to provide richer and nuanced explanation of readmissions [13]. In one of the previous analyses on 30 day hospital readmissions, medical conditions like Chronic Obstructive Pulmonary Disorder (COPD), Total Hip Arthroplasty (THA) etc were used primarily [6]. In this analysis, both parametric and non-parametric statistical models and machine learning techniques like Gradient Boosting, Neural Networks and Decision Trees were used and the primary metric used for performance evaluation was Area Under Curve (AUC). Another study was performed to predict 30-day hospital readmission in COPD patients [1]. This study, apart from the medical factors contributing towards a readmission, incorporated cost and proposed several methods to directly incorporate cost into the prediction.

A few studies focussed on all-cause readmissions where predictive models were built with various types of predictors. These studies included fixed patient attributes such as morbidity burden, maternity and disability etc while most focussed on general patient attributes such as previous acute hospital stays, accumulated days, count of emergency department episodes etc [5]. One of the studies involving prediction of all cause readmissions involved prediction of post discharge death and analysed the deterioration of model

performance in population having multiple admissions per patient [12]. A few studies that used LACE index analysed its poor peformance in some older population [3].

## III. DATA DESCRIPTION

The data for this study has been obtained from various health insurance providers in the USA. The demographics data, medical claims data and the pharmacy claims data are collected from these providers. The fields present in each of the datasets are desribed in the following sections.

*1) Demographics Data:* Demographics data is the statistical data which describes the characteristics of a population such as Age, Gender, Income, Race, Education, Employement etc. For the purpose of this study Gender, Age, Ethnicity and Scheme Type are considered as the demographic features of a person. TABLE I describes each of the above mentioned fields.

| Field | Description |
|---|---|
| Gender | Describes the sex of the person (M - Male or F - Female) |
| Age | Age is a numerical value |
| Ethnicity | Describes the race of the person and takes the values White, Asian, Hispanic and Black |
| Scheme Type | Describes the living area of the person and takes the values Large Central Metro, Large Fringe Metro, Medium Metro, Small Metro, Micropolitan, Noncore |

TABLE I: Demographics Data Description Table

*2) Medical Claims Data:* Medical claims data is the information available in medical billing claims forms filled on behalf of a population. This information is gathered from the medical claims submitted by health care providers to the health insurers. The information obtained from these providers is at an individual claim level and consists of the following fields, Service Start Date, Service End Date, Primary Diagnosis Code, Other Diagnosis Codes, CPT Code. The description for each of these fields is given in TABLE II.

*3) Pharmacy Claims Data:* The pharmacy claims data is obtained from the health insurers. This data contains the information related to the drugs prescribed by a doctor - name of the drug, quantity, prescription date, purchase date, price of the drugs etc. For the purpose of this study, Service Date (purchase date) and the NDC Code (a unique 10-digit numeric identifier assigned to each medication) are considered. The description of these two fields is given in TABLE III.

## IV. DATA PROCESSING

In this study, we have conducted predictive modelling at the episode-level. The target variable is whether a patient was readmitted within 30 days or not after being

| Field | Description |
|---|---|
| Service Start date | The begin date of a medical service |
| Service End date | The end date of a medical service |
| Primary Diagnosis Code | The ICD code for the primary disease |
| Other Diagnosis Codes | The ICD codes for the other diagnosed diseases |
| CPT Code | The CPT code of the procedures undergone during the hospital visit |

TABLE II: Medical Claims Data Description Table

| Field | Description |
|---|---|
| Service Date | The date on which the pharmacy drug was purchased. |
| NDC Code | The NDC code representing the drug |

TABLE III: Pharmacy Claims Data Description Table

discharged. Hence the target variable is a binary variable. The objective of this experiment is to construct predictive models to determine how various factors impact a patients re-hospitalization. The process of finding an episode is explained in the following paragraphs. Let us consider the Medical Claims TABLE IV, Demographics TABLE VIII and Pharmacy Claims Table TABLE VI for explaining the data processing steps.

As the medical claims data is at an individual claim level (which can be observed from TABLE IV) and not at the episode level, certain methods of identifying an admission were followed. An admission can be identified by the inpatient CPT codes (99231-99236, 99224-99226, 99281-99285, 99291-99292) and the discharge can be identified by 99238, 99239, 99217. But the discharge CPT codes are not used often and hence this method could not be used. The length of stay heuristic is followed here which states that, *Two individual claims are grouped together as a part of an episode if the difference between the service end date of the first claim and the service start date of the second claim is less than 10 days* [7].

After the admissions of each user are found, an admission is treated as a readmission if the difference between the previous admission and the current admission is less than or equal to 30 days. If a particular admission is a readmission, it is removed from the list of admissions and added to the readmission list. In our data this grouping resulted in 40,358 admissions of which 1,880 are readmissions resulting in 4.65% readmission rate.

In the medical claims TABLEIV, claim with ClaimID C2 has the inpatient CPT code. Claims C2 and C3 are grouped together as the difference between service end date and start date of the two claims is less than 10 days. Since the difference between the dates for C3 and C4 is greater

| UserID | ClaimID | Service Start Date | Service End Date | Primary Diagnosis Code | Other Diagnosis Codes | CPT Code |
|---|---|---|---|---|---|---|
| User1 | C1 | 2017-04-01 | 2017-04-01 | 682.50 | 786.50 | 99211 |
| User1 | C2 | 2017-05-01 | 2017-05-03 | 70890 | 40201 | 99281 |
| User1 | C3 | 2017-05-04 | 2017-05-08 | 041.12 | 09320 | 61000 |
| User1 | C4 | 2017-05-21 | 2017-06-09 | 186.19 | 00000 | 99231 |
| User1 | C5 | 2017-07-01 | 2017-07-03 | 37234 | 34200 | 99231 |
| User2 | C6 | 2018-01-03 | 2018-01-08 | 78903 | 49001 | 99231 |
| User2 | C7 | 2018-01-03 | 2018-01-15 | 995.29 | 00000 | 43888 |

TABLE IV: Medical Claims Data Table

| UserID | ID | Service Start Date | Service End Date | Re-admission |
|---|---|---|---|---|
| User1 | A1 | 2017-05-01 | 2017-05-08 | YES |
| User1 | A2 | 2017-07-01 | 2017-07-03 | NO |
| User2 | A3 | 2018-01-03 | 2018-01-15 | NO |

TABLE V: Admissions Table

than 10 days and less than 30 days, C4 is considered to be a readmission. Similarly, the claims are grouped for User2 as well. The admissions table after grouping the claims using the above mentioned process is shown in TABLE V.

The predictor variables considered for this study are Comorbidities, Demographics, Length of Stay, Medications during the admission, Number of previous admissions, Number of previous emergency department admissions, Admitting Diagnosis, Number of previous hospital visits, and Admission procedures. All the predictor variables have been derived from the available data and the extraction process of each feature is explained in the following paragraphs. To illustrate the process of extracting each feature, the admissions TABLE V is considered.

| UserID | ClaimID | Service Date | NDC Code |
|---|---|---|---|
| User1 | P1 | 2017-05-05 | 0002759701 |
| User1 | P2 | 2017-05-07 | 5024204062 |
| User2 | P3 | 2018-01-04 | 6057541121 |

TABLE VI: Pharmacy Claims Data Table

*1) Comorbidities:* Comorbidity is the presence of one or more additional diseases co-occurring with a primary disease. For the purpose of this study we considered the following comorbidities - CHF, Valvular, PHTN, PVD, HTN, HTNcx, Paralysis, NeuroOther, Pulmonary, DM, DMcx, Hypothyroid, Renal, Liver, PUD, HIV, Lymphoma, Mets, Tumor, Rheumatic, Coagulopathy, Obesity, Weight-Loss, FluidsLytes, BloodLoss, Anemia, Alcohol, Drugs, Psychoses, Depression. Each of these comorbidities map to a set of ICD codes. Based on the field "Other Diagnosis Codes" in the medical claims data the comorbidities during each admission are identified. For each admission in the TABLE V, the comorbidities present are shown in Table TABLE VII.

| UserID | ID | Comorbidities |
|---|---|---|
| User1 | A1 | CHF, Valvular |
| User1 | A2 | Paralysis |
| User2 | A3 | Pulmonary |

TABLE VII: Comorbidities in each admission

*2) Demographics:* The demographic features considered in this study include Gender, Age Group, Ethnicity, Income Level and Scheme Type. Age group is derived by discretizing the Age into five groups Touch [0-20), Millennials [20-37), GenX [37-49), Boomers [49-68) and Swing (68+). The demographic features of each user in the admissions TABLE V are shown in TABLE VIII.

| UserID | Gender | Age | Ethnicity | Scheme |
|---|---|---|---|---|
| User1 | M | 25 | Asian | Large Central Metro |
| User2 | F | 35 | White | Medium Metro |

TABLE VIII: Demographics Data Table

*3) Length of Stay:* The Length of Stay (LOS) of each admission is derived by subtracting the admission date (which is the service start date of the episode) from the discharge date (which is the service end date of the episode). This a numerical feature and can take values starting from 0. For each admission in the TABLE V, the length of stay is indicated in TABLE IX.

| UserID | ID | LOS |
|---|---|---|
| User1 | A1 | 8 days |
| User1 | A2 | 3 days |
| User2 | A3 | 13 days |

TABLE IX: Length of stay for each admission

*4) Medications:* The NDC codes of the drugs taken during the admission are fetched from the pharmacy claims. These NDC codes are categorized into 100 groups using GPI level 2 categorization. For exampple, a drug with NDC Code 6057541121 will have 60 (first two digits) as its GPI level 2 category. These 100 categories take binary values.For each admission in the TABLE V, the medications taken are shown in TABLE X.

| UserID | ID | Medications |
|--------|-----|-------------|
| User1 | A1 | 00, 50 |
| User1 | A2 | None |
| User2 | A3 | 60 |

TABLE X: Number of previous admissions

*5) Number of Previous Admissions:* Using the derived admissions data from the medical claims data, the number of previous admissions for each admission is derived by taking the count of admissions whose admission date is before the service start date of the current admission. This is a numerical feature and the value is a whole number. For each admission in the TABLE V, the number of previous admissions is shown in TABLE XI.

| UserID | ID | Number of Previous Admissions |
|--------|-----|-------------|
| User1 | A1 | 0 |
| User1 | A2 | 1 |
| User2 | A3 | 0 |

TABLE XI: Number of previous admissions

*6) Number of Previous Emergency Department Admissions:* The emergency department admissions are identified by the set of CPT codes 99281-99285 and these admissions are filtered from the admission data. The number of previous emergency admissions is derived by taking the count of emergency admissions whose admission date is before the service start date of the current admission. This is a numerical feature and the value is a whole number. For each admission in the TABLE V, the number of previous emergency department admissions is shown in TABLE XII.

| UserID | ID | Number of previous ED Admissions |
|--------|-----|-------------|
| User1 | A1 | 0 |
| User1 | A2 | 1 |
| User2 | A3 | 0 |

TABLE XII: Number of previous emergency department admissions

*7) Admitting Diagnosis:* The admitting diagnosis is obtained from the primary diagnosis code of the claim generated on the admission day. Since this is an ICD code and can have upto 70000 values, this field is categorized into 18 body system groups namely:
1) Infectious and parasitic disease

2) Neoplasms
3) Endocrine,nutritional,metabolic,immunity disorders
4) Blood and blood-forming organs
5) Mental disorders
6) Nervous system and sense organs
7) Circulatory system
8) Respiratory system
9) Digestive system
10) Genitourinary system
11) Complications of pregnancy, childbirth and the puerperium
12) Skin and subcutaneous tissue
13) Musculoskeletal system
14) Congenital anomalies
15) Certain conditions originating in the perintal perioud
16) Symptoms, signs and ill-defined conditions
17) Injury and poisoning
18) Factors influencing health status and contact with health services

For each admission in the TABLE V, the admitting diagnosis at CCS level in each admission is shown in TABLE XIII.

| UserID | ID | Admitting Diagnosis |
|--------|-----|-------------|
| User1 | A1 | Others |
| User1 | A2 | Nervous system and sense organs |
| User2 | A3 | Respiratory system |

TABLE XIII: Admitting Diagnosis for each admission

*8) Number of previous hospital visits:* In the medical claims data, hospital visits can be identified by the following CPT Codes 99218-99223 and 99251-99254. The number of previous hospital visits is obtained by taking the count of claims whose service end date is before the service start date of the current admission. For each admission in the TABLE V, the number of previous hospital visits is shown in TABLE XIV.

| UserID | ID | Number of previous hospital visits |
|--------|-----|-------------|
| User1 | A1 | 0 |
| User1 | A2 | 0 |
| User2 | A3 | 0 |

TABLE XIV: Number of previous hospital visits

| UserID | ID | Admission Procedures |
|--------|-----|-------------|
| User1 | A1 | Incision and excision of CNS |
| User1 | A2 | None |
| User2 | A3 | Gastric bypass and volume reduction |

TABLE XV: Admission Procedures for each admission

*9) Admission Procedures:* The procedures taken during an admission are identified through the CPT codes. Using the start and end dates of an admission, the CPT codes

from the medical claims generated between these dates are considered. As the number of CPT codes is huge, they are categorized into 242 groups using the Clinical Classification Software (CCS) [4]. The Clinical Classifications Software (CCS) procedure categorization scheme that can be employed in many types of projects analyzing data on procedures. CCS is based on the International Classification of Diseases, a uniform and standardized coding system. The procedure codes are collapsed into a smaller number of clinically meaningful categories that are sometimes more useful for presenting descriptive statistics than the individual codes. For each admission in the TABLE V, the procedures undertaken are shown in TABLE XV.

The extracted predictor variables from the data are combined together and processed for modelling.

## V. Data Modelling and Results

A key objective of this study was to construct predictive models that can predict whether a patient will be readmitted within 30 days, after being discharged from a hospital unit. After extracting the features all the categorical variables are split into multiple binary columns based on the number of levels. Multiple models with response as *Readmission* were built on the dataset using different prediction tools and their performances were compared. The predictive models are based on Logistic Regression, Principal Component Analysis, Random Forest and Support Vector Machines. The dataset for modelling is split into two sets (80% of the data as training set and 20% as testing set).

*1) Logistic Regression:* Two logistic regression models were built one using all the predictor variables (A) and the other using important variables given by log likelihood feature selection (B). Model A has a Train AUC of 0.716 and Test AUC of 0.663 whereas Model B has a Train AUC of 0.691 and Test AUC of 0.659. The metrics of performance are shown in TABLE XVI and the threshold value for these metrics is $95^{th}$ percentile of the predicted probabilities. These results show that the model with all variables has better performance compared to the model with important variables.

*2) Principal Component Analysis based Regression:* As the modelling ready dataset has very high number of features, Principal Component Analysis(PCA) is performed before and after feature selection on variables. After the PCA, two logistic regression models are built one without feature selection (C) and with feature selection (D). Model C yielded a Train AUC of 0.699 and a Test AUC of 0.655 whereas Model D yielded a Train AUC of 0.684 and a Test AUC of 0.660. The metrics of performance are shown in TABLE XVI and the threshold value for these metrics is $95^{th}$ percentile of the predicted probabilities.

*3) Random Forest Classification:* A grid search is performed using Random Forest Machine Learning technique to predict the chance of readmission. For this, the following values of parameters are considered.
1) Number of trees (ntree): 500, 1000, 150
2) Number of variables in Random Sample at each split (mtry): 20, 30, 40, 50

3) Minimum size of terminal nodes (nodesize): 1, 3, 7, 9
4) Maximum number of terminal nodes the forest can have (maxnodes): 200, 300

Random Forest models using ten fold cross validation are built using all combinations of the above menitioned parameters. The model with the parameters ntree[500], mtry[50], nodesize[7] and maxnodes[300] gave the best performance. The Train AUC of this model is 0.85 and the Test AUC of this model is 0.67. The ROC plots are shown in the figures Fig.1 and Fig.2. The other metrics of performance are shown in TABLE XVI. The threshold for these metrics is the $95^{th}$ percentile of the predicted probabilities. The important features based on gini index from random forest are shown in Fig. 3.
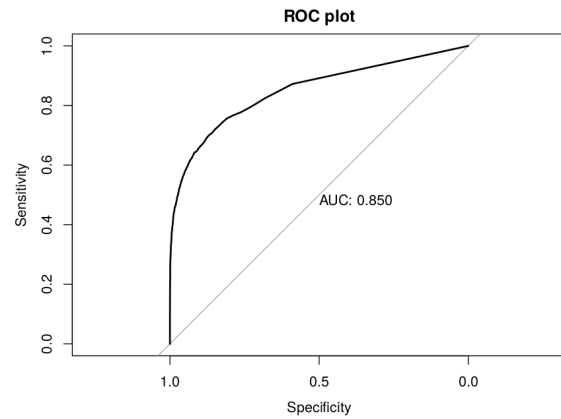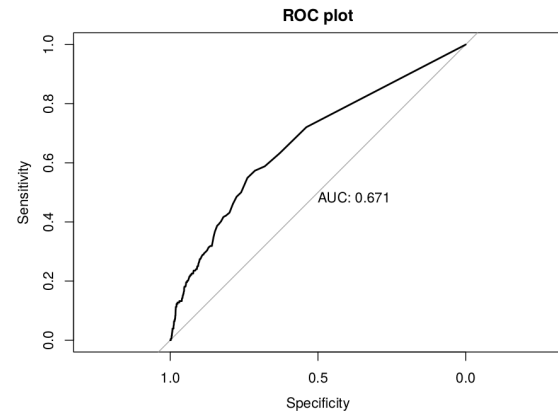


Fig. 1: Train ROC Curve for Random Forest



Fig. 2: Test ROC Curve for Random Forest

*4) Support Vector Machine Classification:* Support Vector Machine (SVM) classfication technique is also used to predict the chance of readmission. Several models using ten fold cross validation are built by tuning the cost of constraints violation (C) parameter with Linear kernels. The values of C are mentioned below. The best SVM model had a Train AUC of 0.66 and Test AUC of 0.64. The other metrics of performance are shown in TABLE XVI. The threshold value for these metrics is $95^{th}$ percentile of the predicted probabilities..
1) Kernel: Linear

| Type | Train AUC | Test AUC | Train Specificity | Test Specificity | Train Sensitivity | Test Sensitivity |
|---|---|---|---|---|---|---|
| Logistic Regression without Feature selection | 0.716 | 0.663 | 0.992 | 0.992 | 0.057 | 0.0591 |
| Logistic Regression with Feature selection | 0.691 | 0.659 | 0.991 | 0.991 | 0.0501 | 0.053 |
| PCA based Regression without Feature selection | 0.699 | 0.655 | 0.991 | 0.991 | 0.0419 | 0.0419 |
| PCA based Regression with Feature selection | 0.684 | 0.660 | 0.991 | 0.991 | 0.0419 | 0.0419 |
| Random Forest | 0.85 | 0.67 | 0.92 | 0.90 | 0.62 | 0.28 |
| SVM | 0.66 | 0.64 | 0.51 | 0.50 | 0.63 | 0.62 |

TABLE XVI: Performance Metrics - The above metrics are based on threshold $= 95^{th}$ percentile of predicted probabilities

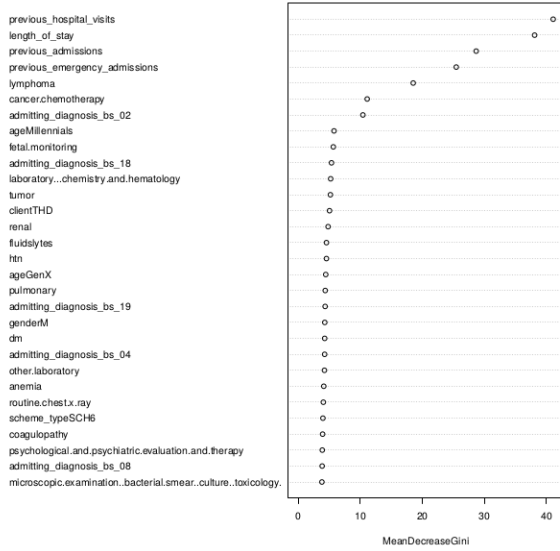2) Cost of constraints violation (C): 0.001, 0.01, 0.05, 0.1, 0.15, 0.2, 0.3, 0.5, 1



Fig. 3: Important Features from Random Forest

## VI. CONCLUSION

Reducing the readmission rate is one of the primary actions that can help in achieving a reduction in healthcare expenses. Different strategies can be implemented using the results from predictive modelling. The ability to recognize patients at high risk of readmission is an important step to improve the quality of care. This also helps in targeting interventions to lower the risk of readmission. In this project, we built models to predict all-cause readmissions using medical claims data. Random Forest classification model had the best AUC value. As a part of future work, we aim to build predictive models focussing on specific medical conditions. Pre-index-admission and Post-index-admission data can be used along with the admission data to understand the crucial causes behind a readmission.

## REFERENCES

[1] C. Baechle, A. Agarwal, R. Behara, and X. Zhu. A cost sensitive approach to predicting 30-day hospital readmission in copd patients. In *2017 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*, pages 317–320, Feb 2017.

[2] Vasiliki Betihavas, Steven A Frost, Phillip J Newton, Peter Macdonald, Simon Stewart, Melinda J Carrington, Yih Kai Chan, and Patricia M Davidson. An absolute risk prediction model to determine unplanned cardiovascular readmissions for adults with chronic heart failure. *Heart, Lung and Circulation*, 24(11):1068–1073, 2015.

[3] Paul E Cotter, Vikas K Bhalla, Stephen J Wallis, and Richard WS Biram. Predicting readmissions: poor performance of the lace index in an older uk population. *Age and ageing*, 41(6):784–789, 2012.

[4] Clinical Classifications Software (CCS) for ICD-9-CM. https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp, 2017.

[5] K. Lemke. A predictive model to identify patients at risk of unplanned 30-day acute care hospital readmission. In *2013 IEEE International Conference on Healthcare Informatics*, pages 551–556, Sept 2013.

[6] R. M. Maddipatla, M. Hadzikadic, D. P. Misra, and L. Yao. 30 day hospital readmission analysis. In *2015 IEEE International Conference on Big Data (Big Data)*, pages 2922–2924, Oct 2015.

[7] Length of Stay Calculation. https://www.ahcancal.org/research_data/trendtracker/Documents/Length%20of%20Stay%20Calculation.pdf, 2014.

[8] Patient Protection and Affordable Care Act. Patient protection and affordable care act. *Public law*, 111(48):759–762, 2010.

[9] Issac Shams, Saeede Ajorlou, and Kai Yang. A predictive analytics approach to reducing 30-day avoidable readmissions among patients with heart failure, acute myocardial infarction, pneumonia, or copd. *Health care management science*, 18(1):19–34, 2015.

[10] Mollie Shulan, Kelly Gao, and Crystal Dea Moore. Predicting 30-day all-cause hospital readmissions. *Health care management science*, 16(2):167–175, 2013.

[11] Douglas S Wakefield and David R Mehr. Risk factors for all-cause hospital readmission within 30 days of hospital discharge. *JCOM*, 20(5), 2013.

[12] Carl Walraven, Jenna Wong, Alan J Forster, and Stephen Hawken. Predicting post-discharge death or readmission: deterioration of model performance in population having multiple admissions per patient. *Journal of evaluation in clinical practice*, 19(6):1012–1018, 2013.

[13] Z. Yu and W. B. Rouse. A deeper look at the causes of hospital readmissions. In *2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, pages 919–923, Dec 2017.