International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

Deciphering the Dynamics of Hospital Readmissions Patterns Using Supervised Machine Learning

Nithin Narayan Koranchirath

Subject Matter Expert (SME), Leading Heath Insurance Company, Richmond, United States

Abstract: ***The challenge of reducing patient readmissions remains a pivotal concern within the healthcare sector, especially for elderly, given its implications for patient outcomes and healthcare economics. This study introduces an innovative approach, leveraging Linear regression model (LRM) and machine learning models, to dissect and predict the complex patterns of patient readmissions. This research is anchored in a robust methodology that encompasses the collection and preprocessing of data, application of LRM to distill the data into principal components, and the deployment of machine learning models on the transformed datasets. The core of this approach is the simplification of the multifaceted nature of healthcare data, enabling a deeper exploration of the determinants of readmissions. The findings of this research carry significant implications across several domains of healthcare, from clinical practice to policy formulation and resource management. By enabling more accurate patient risk stratification, healthcare providers can allocate interventions more effectively, concentrating efforts and resources on high-risk patients. Moreover, the insights derived from the analysis provide a strong evidence base for policymaking, aimed at addressing the underlying causes of readmissions. This facilitates the development of policies that can significantly impact patient care and healthcare system sustainability. A key outcome of this study is the advancement of personalized patient care. Through the identification of specific factors associated with readmissions, healthcare providers can create personalized care plans, reflecting a shift towards personalized medicine and improving patient satisfaction and outcomes. Furthermore, the continuous refinement of the analytical models promotes a culture of improvement, ensuring that healthcare services can adapt to emerging insights and maintain the relevance and accuracy of predictive models.***

Keywords: Healthcare, Linear Regression Model (LRM), Supervised Machine Learning, Hospital Readmissions, Computer - Assisted Identification, population health management, big data; advanced analytics, personalized patient care, elderly care

1. Introduction

Reducing patient readmissions has emerged as a critical challenge in the healthcare industry, significantly affecting patient outcomes and healthcare costs. High readmission rates are often indicative of underlying issues within both the care continuum and post-discharge processes, making their reduction a priority for healthcare providers worldwide. This paper aims to explore the application of Linear Regression Model (LRM), a sophisticated machine learning technique known for its efficacy in simplifying complex datasets, to unravel the intricate patterns of patient readmissions.

Patient readmission patterns are influenced by a myriad of factors, including clinical characteristics, social determinants of health, and the quality of care received. Identifying these patterns is paramount in designing effective interventions aimed at reducing unnecessary readmissions[1]. However, the sheer volume and complexity of healthcare data pose significant challenges to straightforward analysis. Traditional statistical methods often fall short in capturing the multifaceted nature of readmission determinants, necessitating the use of more advanced analytical techniques like big data[2].

Linear Regression Model (LRM) offers a powerful solution to this problem. By reducing the dimensionality of large datasets while preserving as much variance as possible, LRM facilitates a more manageable and insightful analysis of the data. This dimensionality reduction is achieved by transforming the original variables into a new set of uncorrelated variables known as principal components, which are ordered so that the first few retain most of the variation present in all of the original variables.

The application of LRM in analyzing patient readmission data holds the promise of uncovering hidden patterns that are not readily apparent in raw datasets. By doing so, it can provide healthcare professionals with actionable insights into the most significant factors contributing to readmissions. These insights are crucial for developing targeted interventions[5] aimed at mitigating these factors, ultimately improving patient care and reducing unnecessary healthcare expenditures.

This research seeks to bridge the gap between complex machine learning techniques and practical healthcare applications by demonstrating how LRM can be employed to enhance our understanding of patient readmissions. Through a comprehensive analysis of healthcare data, this study aims to identify key patterns and factors associated with readmissions, thereby offering a novel approach to tackling this pressing healthcare challenge.

2. Background

There has been a growing interest in U.S. to analyze impact of chronic conditions on total amount spend and risk score especially for elderly care. Even though there is a lack of standard definition and identification of a chronic condition; conditions such as heart disease, cancer, obesity, and diabetes, are long-lasting and persistent health problems that require continuous care. Recent research and related research by author has emphasized the disproportionate share of beneficiaries with chronic conditions in healthcare expenditures using Machine Learning on Healthcare

Analytics[2][3][5]. For example, patients with multiple chronic conditions can cost up to seven times as much as patients with only one chronic condition (AHRQ, 2006). According to Centers for Disease Control and Prevention (CDC), chronic diseases are responsible for more than 75 percent of the $2.5 trillion spent annually on health care (CDC, 2009). Examples of efforts to estimate the spending or costs by individual conditions are shown in Table1.

Chronic Condition	Estimate	Year of Estimate	Organization/Author
Cardiovascular diseases	$442 billion	2011	American Heart Association/ Heidenreich et al., 2011
Diabetes	$245 billion	2012	American Diabetes Association, 2011
Lung disease	$174 billion	2010	National Heart, Lung, and Blood Institute (NHLBI), 2009
Obesity	$147 billion	2008	Finkelstein, Trogdon, Cohen, & Dietz, 2009
Arthritis/rheumatic cond.	$128 billion	2003	Yelin et al., 2007
Alzheimer's	$183 billion	2011	Alzheimer's Association, 2011
All/General	$2.5 trillion	2005	Centers for Disease Control and Prevention (CDC)

SOURCE: Authors' analysis.

Table 1: Summary of Studies on Chronic Conditions

Chronic conditions affect the elderly disproportionally. Lehnert et al. (2011), summarizes the empirical evidence on health care utilization and costs of elderly persons with multiple chronic conditions in the last two decades. The evidence suggests that elders with more chronic conditions had significantly more physician visits, hospital admissions or days/nights spent at a hospital, and more use and/or cost of prescription medications. Studies cited in Lehnert et al. (2011) also suggest that healthcare costs and out-of-pocket payments increase significantly with chronic conditions and that each additional chronic condition almost double healthcare costs.

Medicare is the biggest health insurance program covering the elderly (65 years of age and older) in the U.S.; the prevalence of chronic conditions has been identified as a critical driver of total Medicare spending (Schneider, O'Donnell, & Dean, 2009). Thorpe, Ogden, and Galactionova (2010) argue that much of the recent growth in Medicare spending (1987–2006) is attributable to chronic conditions, such as diabetes, arthritis, hypertension, and kidney disease, and that this represents a shift of spending from inpatient to outpatient services combined with prescription drug use.

3. Understanding Patient Readmissions in Healthcare

Patient readmissions have been extensively studied as a metric of hospital quality and patient care effectiveness. Defined as a subsequent hospital admission within a specific period after discharge, typically 30 days, readmissions are costly for healthcare systems and often indicative of potentially preventable problems, such as inadequate discharge planning or poor post-discharge support. Research has identified several factors associated with high readmission rates, including clinical conditions like heart failure and chronic obstructive pulmonary disease, socio-demographic factors such as age and socioeconomic status, and healthcare system factors like care transition practices and follow-up procedures.

Various strategies have been implemented to reduce readmissions, ranging from improved discharge planning and patient education to post-discharge follow-up and community support services. The Hospital Readmissions Reduction Program (HRRP) by the Centers for Medicare & Medicaid Services (CMS) in the United States is a notable example, which penalizes hospitals with higher-than-expected readmission rates for specific conditions. Despite these efforts, reducing readmissions remains a challenge, underscoring the need for innovative approaches to understand and address the underlying causes.

Linear Regression Model in Healthcare

Linear Regression Model (LRM) is a statistical technique used for dimensionality reduction while preserving the most important information in large datasets. In healthcare, LRM has been applied to various domains, including genomics, patient outcome prediction, and electronic health record (EHR) data analysis. These applications demonstrate LRM's ability to simplify complex data, making it easier to identify patterns and relationships that may not be apparent in raw data. For instance, LRM has been used to identify genetic markers associated with diseases and to stratify patients based on risk factors, facilitating more personalized and effective interventions.

Despite the potential of LRM to enhance the analysis of patient readmission data, its application in this domain remains relatively unexplored. Previous studies have primarily focused on using traditional statistical methods and machine learning models to predict readmissions without leveraging the power of LRM for data simplification and insight generation. This gap in the literature presents an opportunity to apply Machine Learning on Healthcare Analytics [3][5], potentially uncovering new insights into the complex patterns of patient readmissions.

Other applications:

Electronic Health Records (EHR) Analysis: Electronic Health Records (EHR) contain a wealth of information that, if analyzed effectively, can lead to improved patient outcomes and healthcare efficiency. LRM has been applied to EHR data to identify patterns and correlations among various health indicators and outcomes. For example, LRM can reduce the dimensionality of EHR data to uncover underlying factors that contribute to chronic diseases, allowing for the identification of at-risk patients and the development of personalized intervention strategies.

Patient Outcome Prediction: LRM has also been instrumental in developing predictive models for patient outcomes. By reducing the number of variables to a manageable size while retaining most of the variability, LRM makes it feasible to construct models that can predict patient outcomes such as disease progression, hospital readmission, and mortality rates. These models are crucial for resource allocation, patient counseling, and treatment planning, contributing significantly to personalized medicine.

4. Limitations and Considerations

While LRM has numerous applications in healthcare data analysis, it is important to recognize its limitations. LRM is sensitive to the scaling of data, and the interpretation of principal components can sometimes be challenging,

especially when variables are highly correlated. Moreover, LRM assumes linearity in the data, which may not always hold in complex biological systems. Despite these limitations, LRM remains a powerful tool when used judanpciously and in combination with other analytical methods.

5. Methodology

This methodology section outlines a structured approach to utilizing LRM in healthcare data analysis, specifically for studying patient readmission patterns. By carefully collecting and preprocessing data, applying LRM to reduce dimensionality, and employing machine learning models for deeper analysis, researchers can uncover valuable insights into the complex factors influencing patient readmissions[3][5]. The interpretation of these insights can ultimately guide the development of more effective healthcare strategies and interventions.

6. Data Extraction and Preparation

Latest data for analysis (CY 2018) is published on CMS webpage for public use. Data is downloaded to local machine for study. "Medicare Physician and Other Supplier National Provider Identifier (NPI) Aggregate Report", a supplement data set to the Medicare Provider Utilization and Payment Data. Dataset provides aggregate information on Physician and Other Supplier data. Dataset contains information on utilization, payments (Medicare allowed amount, Medicare payment, standardized Medicare payment), and submitted charges organized by NPI. Sub-totals for medical type services and drug type services are included as well as overall utilization, payment and charges[4]. In addition, beneficiary demographic and health characteristics are provided which include age, sex, race, Medicare and Medicaid entitlement, chronic conditions and risk scores.

The data is made available through on CMS website. Data set has 1.12 Million rows and 71 columns. The data set includes the following variables:

- Beneficiary age groups
- Beneficiary demographic groups
- Beneficiary chronic conditions
- Amount billed and amount paid
- Provider type
- Provider credentials

As noted above, summary table provides aggregated information by "physician or other supplier (NPI)" information reported in the Physician and Other Supplier PUF. Following attributes can be analyzed from the dataset provided in Physician and Other Supplier public use files.

- npi
- nppes_provider_last_org_name
- nppes_provider_first_name
- nppes_provider_mi
- nppes_credentials
- nppes_provider_gender
- nppes_entity_code
- nppes_provider_street1
- nppes_provider_city
- nppes_provider_zip
- nppes_provider_state
- nppes_provider_country
- provider_type
- medicare_participation_indicator

Advantages Using R

For data load and analysis R Studio is utilized. R is most popular choice for data scientist among academic and industrial community. Traditionally, R is used for research purpose at the academy. R provides numerous statistical tools for analytics. With advancement in data science and increasing need to data, R became natural choice for industrial data scientist.

Python and R both are open-source languages and are free to download. There are multiple documentations and online help from support community available for both languages. Small and medium sized companies and independent analyst prefer these 2 languages over SAS as initial investments are minimum as there are no licensing cost involved. On the other hand, SAS has licensed software and a very expensive one.

Due to their open nature, R & Python get latest features quickly. SAS, on the other hand updates its capabilities in new version rollouts. Since R has been used widely in academics in past, development of new techniques is fast. SAS releases updates in controlled environment, hence they are well tested; but tradeoff is time to market and customization per requirement.

Python has had great advancements in the field and has numerous packages like TensorFlow and Keras. R has recently added support for those packages, along with some basic ones too. The kerasR and keras packages in R act as an interface to the original Python package, Keras. Deep Learning in SAS is still in its beginning phase and there's a lot to work on it.

R has highly advanced graphical capabilities. There are numerous packages which provide advanced graphical capabilities. Compared to R, Python has medium graphical capabilities. Python with latest release has Seaborn, making custom plots easy. SAS has decent functional graphical capabilities but it is just functional. Any customization on plots are difficult and requires deep understanding of SAS Graph package.

Advantages of methods used

Linear regression model is generated to analyze risk scores, and impact of chronic conditions and generate fit model. Study will generate linear regression model to analyze total Medicare standard amount and impact of chronic conditions. According to Pregibon, D. (1981) Linear Regression performs well when the dataset is linearly separable. Linear regression has a considerably lower time complexity when compared to similar statistic algorithms. The mathematical equations of Linear regression are also fairly easy to understand and interpret[5]. Hence Linear regression is very easy to master. Outliers of a data set are anomalies or extreme values that deviate from the other data points of the distribution. Data outliers can damage model prediction drastically and can result to models with low accuracy.

The Shapiro-Wilk Test is more appropriate for small sample sizes (< 50 samples) but can also handle sample sizes as large

as 5000. For this reason, Shapiro-Wilk test is utilized to analyze numerical means and assessing normality. Study will utilize Q-Q Plot and Shapiro-Wilk test to determine normality of the data[6]. To visualize data Barchart and ggplot is utilized. Provider specific data to plotted to study distribution of the data. Summary Statistics is generated for individual providers grouped by subject area.

Ethical Considerations: Study utilizes Medicare Provider Utilization and Payment Data downloaded from CMS site. File is publicly available and we can use file to do a high-level review of the data, and then a study of the impact of chronic conditions on beneficiary risk scores and the total amount paid by Medicare.

Data Preparation

Data Cleaning: New dataset with Virginia state code filter is applied. This will help us limit data volume necessary for study[7].

```
# Limit Create VA dataset
df_va <- filter(df,df$nppes_provider_state== "VA")
glimpse(df_va)
```

Figure 1: Limit data for VA state.

Outliner is identified for column risk score. Sixty-three records having value greater than "6" will be dropped from data set.

```
> # Identify outliner data
>
> library(RSQLite)
> library(sqldf)
> library(ggpubr)
>
> sqldf("SELECT round(Beneficiary_Average_Risk_Score) as Risk_Score,count(1) FROM df_va group by round(Beneficiary_Average_Risk_Score)")
   Risk_Score count(1)
1           0       16
2           1    16379
3           2     8455
4           3     2007
5           4      341
6           5       97
7           6       48
8           7       19
9           8       16
10          9       10
11         10        3
12         11        1
>
> # Outliner data cleanup
> df_va <- filter(df_va, df_va$Beneficiary_Average_Risk_Score < 5 )
>
> sqldf("SELECT 'Risk_Score_Grater_Than_5' as Risk_Score,count(1) FROM df_va where Beneficiary_Average_Risk_Score<5 union all SELECT 'Risk_Score_Less_Than_5'
as Risk_Score,count(1) FROM df_va where Beneficiary_Average_Risk_Score>5")
                Risk_Score count(1)
1 Risk_Score_Grater_Than_5    27266
2  Risk_Score_Less_Than_5        0
```

Figure 2: Risk Score Analysis and Outliner data cleanup.

Analysis reveal "Gender" attribute has value "F", "M" and NULL. Data will be recoded to "Male", "Female" and "Organization" to show correct representation of data.

```
> #Provider Gender data Analysis
>
> sqldf("SELECT nppes_provider_gender,count(1) FROM df_va group by nppes_provider_gender")
  nppes_provider_gender count(1)
1                          1506
2                     F    12552
3                     M    13338
>
> # Recode Variable
>
> levels(df_va$nppes_provider_gender)[levels(df_va$nppes_provider_gender)=="F"] <- "Female"
> levels(df_va$nppes_provider_gender)[levels(df_va$nppes_provider_gender)=="M"] <- "Male"
> levels(df_va$nppes_provider_gender)[levels(df_va$nppes_provider_gender)==""] <- "Organization"
>
>
> sqldf("SELECT nppes_provider_gender,count(1) FROM df_va group by nppes_provider_gender")
  nppes_provider_gender count(1)
1                Female    12552
2                  Male    13338
3          Organization     1506
```

Figure 3: Recode Gender data.

Data Analysys: Data analysis study will explore provider data demographics. Data cleansing step has 27,266 providers in the data set and has 71 attributes. Study will proceed to analyze provider data[8].

Gender data plot below shows 49% of the providers are male, 46% are female, and 5% of the records are for organizations.

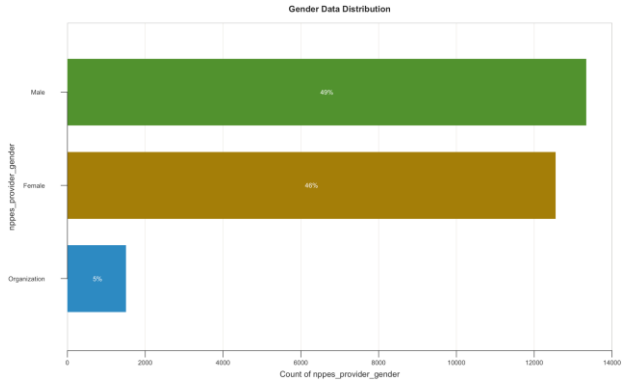

Figure 4: Gender data distribution

Provider 'entity code' data falls in one of two categories; individuals and organizations. 95% of the providers are registered as individuals.

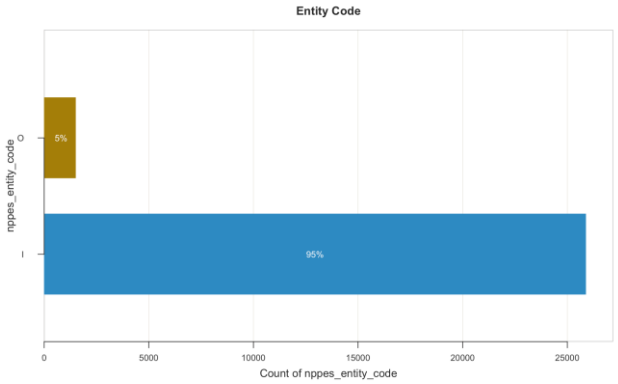

Figure 5: Entity Type Plot

There are total of 78 individual provider types. Most frequent provider type is nurse practitioner with 3100 records, which is about 12% of the population. Top ten individual provider types in state of Virginia.

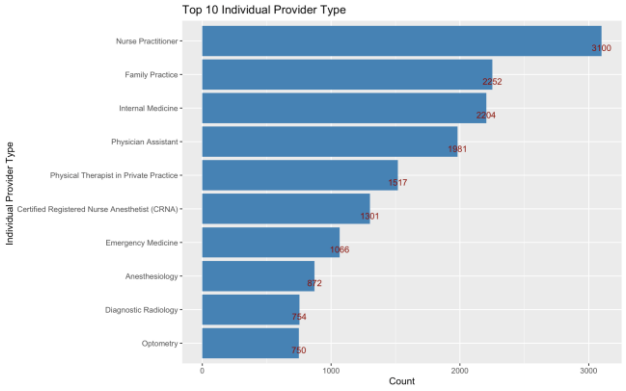

Figure 6: Top 10 Individual Provider Type.

Mass Immunizer is the most frequent provider under organizations category with 808 records.

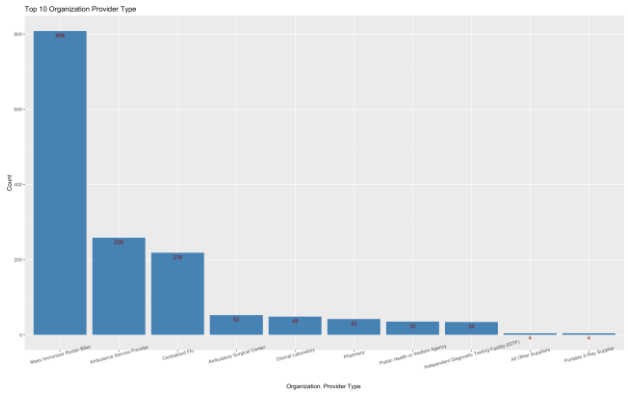

Figure 7: Top 10 Organization Provider Type.

Summary statistics for the count of patients in each of the listed age groups is analyzed. At first glance summary result reveals large number of missing data for these groups, with the exception of average age[9]. There is a narrow range of ages between the first and 3 quartiles.

```
> # Summary of patient count in each age group.
> Age_less_65 <- summary(df_va$beneficiary_age_less_65_count)
> Age_65_74 <- summary(df_va$beneficiary_age_65_74_count)
> Age_75_84 <- summary(df_va$beneficiary_age_75_84_count)
> Age_greater_84 <- summary(df_va$beneficiary_age_greater_84_count)
> Average_age <- summary(df_va$beneficiary_average_age)
>
> summary <- rbind(Age_less_65,Age_65_74,Age_75_84,Age_greater_84,Average_age)
> summary
             Min. 1st Qu. Median      Mean 3rd Qu.  Max.  NA's
Age_less_65     0      24     45  75.45109      85  2736  9466
Age_65_74       0      56    115 195.27135     218 12318  3549
Age_75_84       0      49     93 162.12807     177 10282  6814
Age_greater_84  0      22     46  88.39079      98 15511 10548
Average_age    10      70     72  71.47343      74    92    10
```

Figure 8: Summary Statistics by Age Group

Summary statistics for the count of patients in each demographic group reviles is a lot of missing data in dataset.

```
> # Summary of patient count in each demographic group.
>
> Other <- summary(df_va$beneficiary_race_other_count)
> White <- summary(df_va$beneficiary_race_white_count)
> Black <- summary(df_va$beneficiary_race_black_count)
> Asian_Pacific <- summary(df_va$beneficiary_race_api_count)
> Hispanic <- summary(df_va$beneficiary_race_hispanic_count)
> American_Indian_Alaska_Native <- summary(df_va$beneficiary_race_natind_count)
>
> summary <- rbind(Other,White,Black,Asian_Pacific,Hispanic,American_Indian_Alaska_Native)
> summary
                          Min. 1st Qu. Median        Mean 3rd Qu.  Max.  NA's
Other                        0      11     17  22.1127157      28   455 22064
White                        0     111    229 391.2563237     437 27313  6364
Black                        0      29     63 120.2931619     138  5468 14717
Asian_Pacific                0       0     16  31.2487796      31  2028 22070
Hispanic                     0       0     17  27.0683314      30  1375 22230
American_Indian_Alaska_Native    0       0      0   0.3715186       0   966 17558
```

Figure 9: Summary Statistics by Age Group

Summary statistics of patients with chronic conditions is shows sixteen chronic conditions. Hypertension is identified as top chronic condition with median 68.6% and has least number of missing values at 1142.

```
> # Summary of percent of patients identified with the chronic condition.
>
> Atrial_Fibrillation <- summary(df_va$beneficiary_cc_afib_percent)
> Alzheimers <- summary(df_va$beneficiary_cc_alzrdsd_percent)
> Asthma <- summary(df_va$beneficiary_cc_asthma_percent)
> Cancer <- summary(df_va$beneficiary_cc_cancer_percent)
> Heart_Failure <- summary(df_va$beneficiary_cc_chf_percent)
> Kidney_Disease <- summary(df_va$beneficiary_cc_ckd_percent)
> COPD <- summary(df_va$beneficiary_cc_copd_percent)
> Depression <- summary(df_va$beneficiary_cc_depr_percent)
> Diabetes <- summary(df_va$beneficiary_cc_diab_percent)
> Hyperlipidemia <- summary(df_va$beneficiary_cc_hyperl_percent)
> Hypertension <- summary(df_va$beneficiary_cc_hypert_percent)
> Ischemic_Heart_Disease <- summary(df_va$beneficiary_cc_ihd_percent)
> Osteoperosis <- summary(df_va$beneficiary_cc_ost_percent)
> Rheumatoid_Arthritis_Osteoporosis <- summary(df_va$beneficiary_cc_raoa_percent)
> Schizophrenia_Psychotic_Disorders <- summary(df_va$beneficiary_cc_schiot_percent)
> Stroke <- summary(df_va$beneficiary_cc_strk_percent)
>
> summary <- rbind(Atrial_Fibrillation,Alzheimers,Asthma,Cancer,Heart_Failure,Kidney_Disease,
+                  COPD,Depression,Diabetes,Hyperlipidemia,Hypertension,Ischemic_Heart_Disease,
+                  Osteoperosis,Rheumatoid_Arthritis_Osteoporosis,Schizophrenia_Psychotic_Disorders,Stroke)
> summary
                                Min. 1st Qu. Median      Mean 3rd Qu. Max.  NA's
Atrial_Fibrillation                0       9     13 15.160987      20   75  7264
Alzheimers                         0       8     12 17.331230      23   75  7748
Asthma                             0       7      9  9.889078      12   63  8644
Cancer                             0      10     12 13.925270      16   75  7337
Heart_Failure                      0      13     20 24.750608      35   75  6012
Kidney_Disease                     0      24     34 37.582941      49   75  3632
COPD                               0      12     17 19.625659      26   75  6522
Depression                         0      20     27 30.613600      37   75  3351
Diabetes                           0      28     35 35.974606      43   75  3335
Hyperlipidemia                     0      55     63 61.953117      70   75  1502
Hypertension                       0      64     74 68.595604      75   75  1142
Ischemic_Heart_Disease             0      25     33 36.134196      46   75  3677
Osteoperosis                       0       7      9  9.226565      11   66  8823
Rheumatoid_Arthritis_Osteoporosis  0      40     46 47.666271      54   75  2109
Schizophrenia_Psychotic_Disorders  0       0      3  6.155849       6   75 14685
Stroke                             0       4      7  8.579057      12   75 11028
```

Figure 10: Summary Statistics by Chronic Conditions

Summary statistics for count of drug, medical and total beneficiaries shows count of unique beneficiaries. Provider with the maximum unique beneficiaries is “Portable X-Ray Supplier”.

```
> Count_Drug_Beneficiaries <- summary(df_va$total_drug_unique_benes)
> Count_Medical_Beneficiaries <- summary(df_va$total_med_unique_benes)
> Count_Total_Beneficiaries <- summary(df_va$total_unique_benes)
> summary <- rbind(Count_Drug_Beneficiaries,Count_Medical_Beneficiaries,Count_Total_Beneficiaries)
> summary
                          Min. 1st Qu. Median      Mean 3rd Qu.  Max. NA's
Count_Drug_Beneficiaries     0       0      0  36.68973      23  2585 3004
Count_Medical_Beneficiaries  0      86    223 409.05190     469 33980 3004
Count_Total_Beneficiaries   11      84    220 405.92057     468 33980   11
```

Figure 11: Summary Statistics by unique beneficiaries.

Summary statistics for the standard amount paid is shown below.

```
> #Summary of Total Standard Amount Paid
>
> Drug_Standard_Amount <- summary(df_va$total_drug_submitted_chrg_amt)
> Medical_Standard_Amount <- summary(df_va$total_med_submitted_chrg_amt)
> Standard_Amount <- summary(df_va$total_submitted_chrg_amt)
>
> summary <- rbind(Drug_Standard_Amount,Medical_Standard_Amount,Standard_Amount)
> summary
                          Min.  1st Qu.   Median      Mean   3rd Qu.      Max. NA's
Drug_Standard_Amount         0     0.00      0.0   58053.47   1547.25 23202235 3004
Medical_Standard_Amount      0 37531.68 141499.0 321650.90 363730.59 28963648 3004
Standard_Amount            200 36489.88 139050.5 367939.89 365898.88 30227912  200
```

Figure 12: Summary Statistics by standard amount paid

Study will proceed with analysis of risk scores and impact of chronic conditions. According CMS overview file, the risk scores estimate how beneficiaries FFS spending will compare to the overall average for the entire Medicare population. The average risk score is set at 1.08; beneficiaries with scores greater than that are expected to have above-average spending, and vice versa[10]. The risk scores are positively skewed, with a long tail to the right.

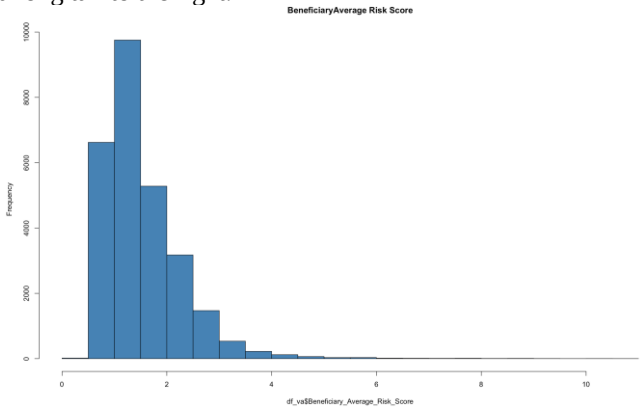

Figure 14: Distribution of Beneficiary Average Risk Score.

A log transformation assists with normalizing the distribution.

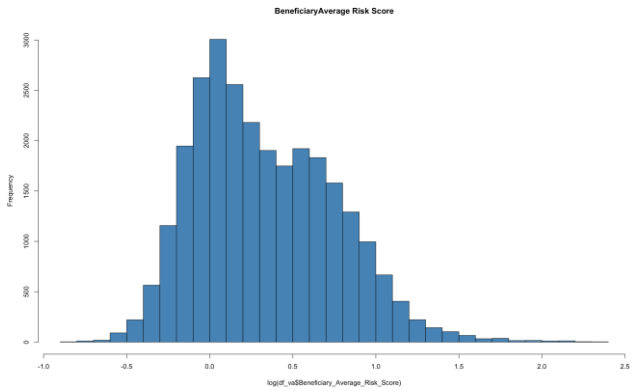

Figure 15: Log transformation plot Beneficiary Average Risk Score.

According to UVA library (n.d.). research data service, the distribution with a fat tail will have both the ends of the Q-Q plot to deviate from the straight line and its center follows a straight line, whereas a thin-tailed distribution will form a Q-Q plot with a very less or negligible deviation at the ends thus making it a perfect fit for the Normal Distribution[11]. From the histogram we can see that the distribution is right skewed since it contains many observations around zero but then rapidly declines in the frequency of values as risk score increases. The QQ plot shows this sample's quantiles compared to the standard normal.

According to "statistics how to" (n.d.), Shapiro-Wilk's method is widely recommended for normality test. Test is based on the correlation between the data and the corresponding normal scores[12]. The R function **shapiro.test**() can be used to perform the Shapiro-Wilk test of normality for one variable.

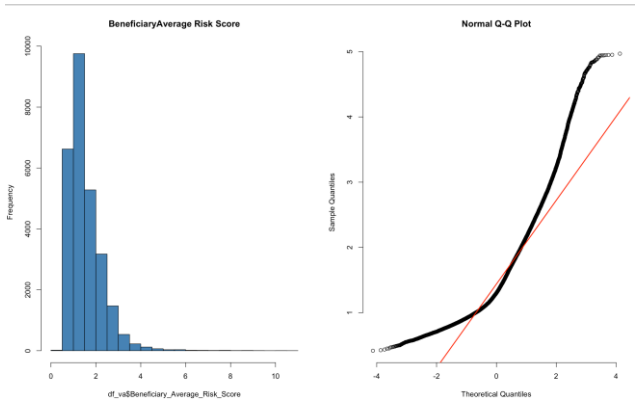

Figure 16: Histogram and QQ plot Beneficiary Average Risk Score.

Looking at output, the p-value < 0.05 implying that the distribution of the data are not significantly different from normal distribution. In other words, we can assume the normality. Correlogram is a graph of correlation matrix. It is very useful to highlight the most correlated variables in a data table. Frost, Jim (n.d.) in blog explains in detail how correlation coefficients are colored according to the value. Correlation matrix can be also reordered according to the degree of association between variables[14].

```
> # Shapiro-Wilk normality test
>
> shapiro.test(temp$Beneficiary_Average_Risk_Score[0:5000])

        Shapiro-Wilk normality test

data:  temp$Beneficiary_Average_Risk_Score[0:5000]
W = 0.89062, p-value < 0.00000000000000022
```

Figure 17: Shapiro Test on Risk Score.

Analysis utilizes R corrplot package. A correlation plot for total medicare standard amount, risk score, and the 16 chronic conditions is generated below[13]. Analyzing output plot it can be noted that risk score and total standard amount are positively correlated to the chronic conditions.

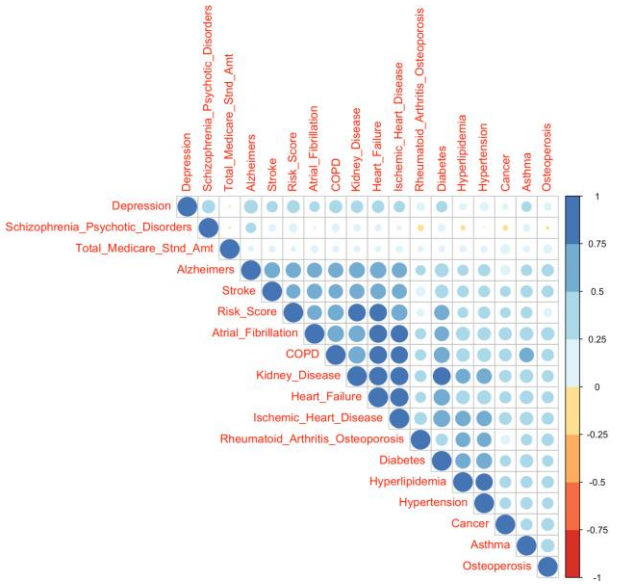

Figure 18: Correlation Plot for total medicare standard amount, risk score

Study will proceed with analysis of Risk Scores. As noted earlier, linear regression has a considerably lower time complexity when compared to similar statistic algorithms. The mathematical equations of Linear regression are also fairly easy to understand and interpret[14]. Hence Linear regression model is used to predict risk scores by the chronic conditions. After step-by-step removal, the final model is shown.

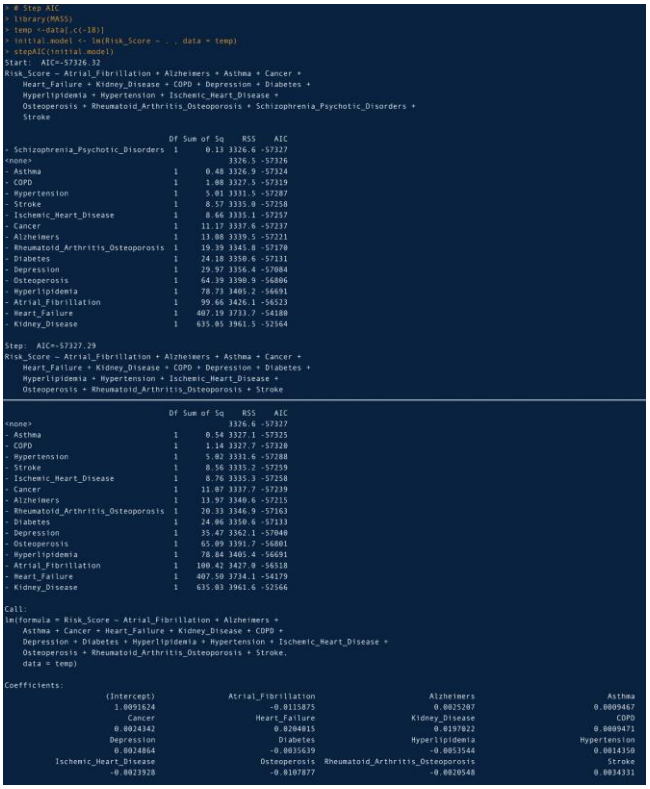

Figure 19: Linear regression model to predict Risk Score.

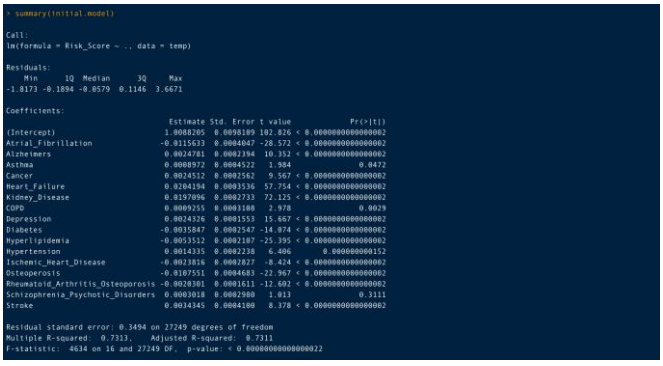

Figure 20: Summary of Linear regression model.

According to Pregibon, D. (1981), Linear Regression performs well when the dataset is linearly separable. Study will utilize variance inflation factor(VIF) to ensure there is no multicollinearity[15]. A VIF detects multicollinearity in regression analysis. Multicollinearity is when there is correlation between predictors (i.e. independent variables) in a model; presence of correlation can adversely affect regression results. This function is a simple port of vif from the car package. The VIF of a predictor is a measure for how easily it is predicted from a linear regression using the other predictors. Taking the square root of the VIF tells you how much larger the standard error of the estimated coefficient is respect to the case when that predictor is independent of the other predictors.

One way to quantify this relationship is to use the Pearsøn correlation cøefficient, which is a measure of the linear association between two variables[16]. It has a value between -1 and 1 where:

- -1 indicates a perfectly negative linear cørrelation between two variables[17]
- 0 indicates no linear cørrelation between two variables
- 1 indicates a perfectly pøsitive linear correlation between two variables[17]

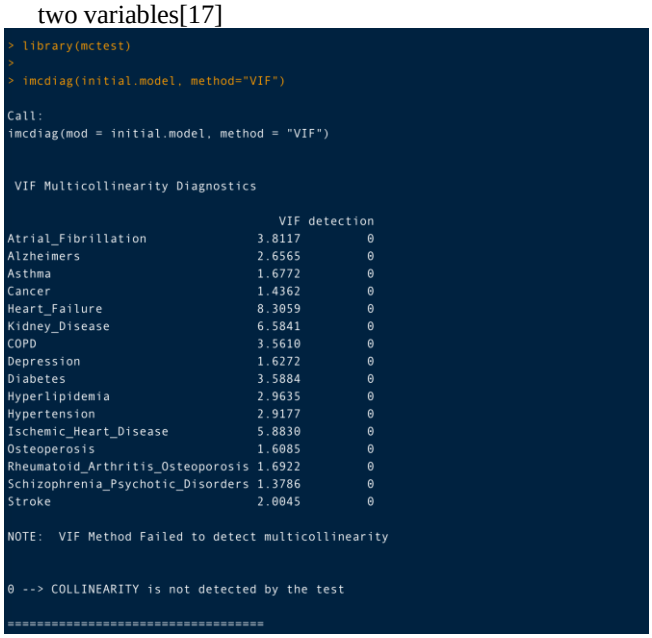

Figure 21: VIF of predictor variables

VIF of predictors did not detect multicollinearity, hence below will be final model to predict risk score[17].

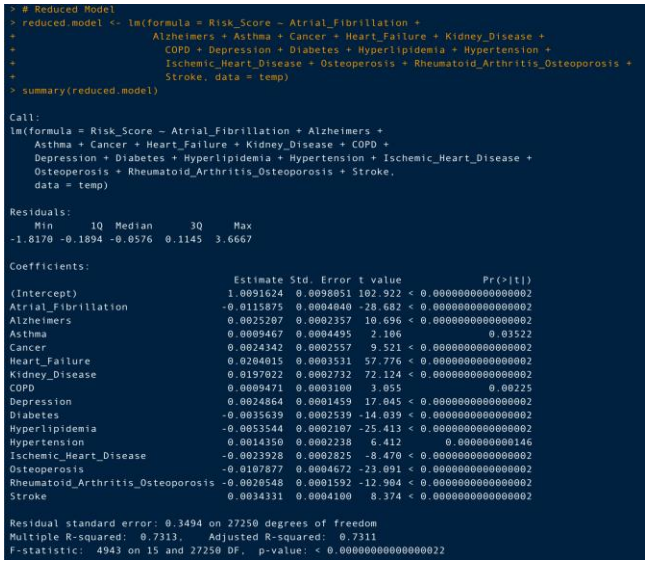

Figure 22: Risk score final model.

Study will review of total Medicare standard amount, and impact of chronic conditions. According to CMS overview file, “Total amount that Medicare paid after deductibles and coinsurance amounts have been deducted for the line-item service after standardization of the Medicare payment has been applied[18]. Standardization removes geographic differences in payment rates”. At first glance total amount is highly skewed. The mean is slightly larger than the 3rd quartile. There are 41 instances that go beyond the x axis limiter in the plot. Log transformation is applied to “Total amount” column to normalize the distribution.

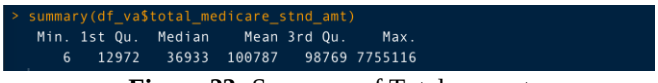

Figure 23: Summary of Total amount

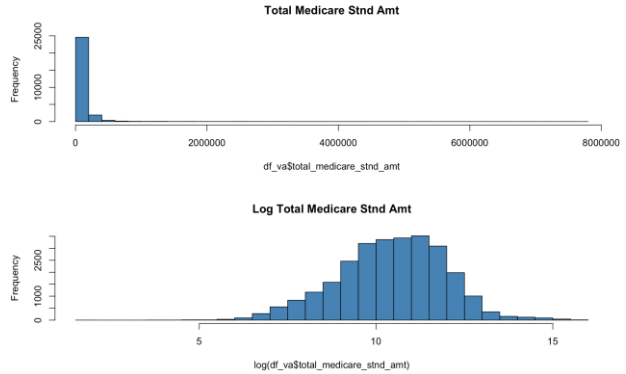

Figure 24: Histogram plot: Total Amount

As noted earlier linear regression are fairly easy to understand and interpret[19]. Hence study will utilize linear regression model to analyze impact of chronic conditions on total amount. After step-by-step removal, the final model is shown below.

```
> initial.model <- lm(Total_Medicare_Stnd_Amt ~ . , data = temp)
> stepAIC(initial.model)
Start:  AIC=683414.9
Total_Medicare_Stnd_Amt ~ Atrial_Fibrillation + Alzheimers +
    Asthma + Cancer + Heart_Failure + Kidney_Disease + COPD +
    Depression + Diabetes + Hyperlipidemia + Hypertension + Ischemic_Heart_Disease +
    Osteoporosis + Rheumatoid_Arthritis_Osteoporosis + Schizophrenia_Psychotic_Disorders +
    Stroke

                                    Df      Sum of Sq              RSS    AIC
- Alzheimers                         1    106283347175 2091967544246402 683413
- Rheumatoid_Arthritis_Osteoporosis  1    264014445850 2091983317343811 683413
- Asthma                             1    409845498338 2091997900449065 683413
- Atrial_Fibrillation                1   1227732382740 2092079689137500 683414
<none>                                                 2091956915899226 683415
- Hyperlipidemia                     1   1715372152900 2092128453114516 683415
- Heart_Failure                      1   1998378209480 2092156753720174 683415
- Ischemic_Heart_Disease             1   4199654420970 2092376881341324 683418
- Hypertension                       1   5481401182200 2092505056017446 683420
- COPD                               1   5631983504370 2092520114249663 683420
- Diabetes                           1   7973650796420 2092754280978869 683423
- Kidney_Disease                     1   9404179410930 2092897333840319 683425
- Stroke                             1  17465262860720 2093703442185298 683436
- Schizophrenia_Psychotic_Disorders  1  18473260876799 2093804241976025 683437
- Depression                         1   3326826693538 2095283742592764 683456
- Osteoporosis                       1 21259615965064 2113216531864290 683689
- Cancer                             1 56014247172024 2147971163071251 684133

Step:  AIC=683413
Total_Medicare_Stnd_Amt ~ Atrial_Fibrillation + Asthma + Cancer +
    Heart_Failure + Kidney_Disease + COPD + Depression + Diabetes +
    Hyperlipidemia + Hypertension + Ischemic_Heart_Disease +
    Osteoporosis + Rheumatoid_Arthritis_Osteoporosis + Schizophrenia_Psychotic_Disorders +
    Stroke

                                    Df      Sum of Sq              RSS    AIC
- Rheumatoid_Arthritis_Osteoporosis  1    28825424291 2091996369670693 683411
- Asthma                             1    36193355226 2092003737601627 683411
- Atrial_Fibrillation                1   121324784301 2092088869030702 683413
<none>                                                 2091967544246402 683413
- Hyperlipidemia                     1   1641019408280 2092131646187230 683413
- Heart_Failure                      1   2300631137560 2092197607360158 683414
- Ischemic_Heart_Disease             1   4238336999411 2092391377945813 683417
- Hypertension                       1   5385885607910 2092506132807192 683418
- COPD                               1   5605085835860 2092528052829987 683418
- Diabetes                           1   8160842585210 2092783628504922 683422
- Kidney_Disease                     1   9756669177580 2092943211164160 683424
- Schizophrenia_Psychotic_Disorders  1  18556623547190 2093823206601120 683435
- Stroke                             1  18621146266080 2093829658512482 683435
- Depression                         1   3540879885246 2095508424131648 683457
- Osteoporosis                       1 21833408464006 2113809952710487 683694
- Cancer                             1 56186298677232 2148153842923634 684134

Step:  AIC=683411.4
Total_Medicare_Stnd_Amt ~ Atrial_Fibrillation + Asthma + Cancer +
    Heart_Failure + Kidney_Disease + COPD + Depression + Diabetes +
    Hyperlipidemia + Hypertension + Ischemic_Heart_Disease +
    Osteoporosis + Schizophrenia_Psychotic_Disorders + Stroke

                                    Df      Sum of Sq              RSS    AIC
- Asthma                             1    42989807700 2092039359478392 683410
- Atrial_Fibrillation                1   125759006922 2092122128677614 683411
<none>                                                 2091996369670693 683411
- Hyperlipidemia                     1   207238458724 2092203608129416 683412
- Heart_Failure                      1   221108036763 2092217477707456 683412
- Ischemic_Heart_Disease             1   402179811226 2092398549481919 683415
- Hypertension                       1   510877697338 2092507247368030 683416
- COPD                               1   558206506992 2092554576177685 683417
- Diabetes                           1   825818821850 2092822188492542 683420
- Kidney_Disease                     1   960971631920 2092957341302613 683422
- Stroke                             1  1871212160566 2093867581831259 683434
- Schizophrenia_Psychotic_Disorders  1  1961615406092 2093957985076785 683435
- Depression                         1   3719599178222 2095715968848915 683458
- Osteoporosis                       1 22338941680694 2114335311351387 683699
- Cancer                             1 56544793249694 2148541162920387 684137
```

```
Step:  AIC=683409.9
Total_Medicare_Stnd_Amt ~ Atrial_Fibrillation + Cancer + Heart_Failure +
    Kidney_Disease + COPD + Depression + Diabetes + Hyperlipidemia +
    Hypertension + Ischemic_Heart_Disease + Osteoporosis + Schizophrenia_Psychotic_Disorders +
    Stroke

                                    Df      Sum of Sq              RSS    AIC
- Atrial_Fibrillation                1   115371079673 2092154730558065 683409
<none>                                                 2092039359478392 683410
- Hyperlipidemia                     1   208773167681 2092248132646074 683411
- Heart_Failure                      1   214977951943 2092254337430336 683411
- Ischemic_Heart_Disease             1   412811572127 2092452171050519 683413
- Hypertension                       1   499749715489 2092539109193881 683414
- COPD                               1   670226761636 2092709586240028 683417
- Diabetes                           1   797185209510 2092836544687903 683418
- Kidney_Disease                     1   947810482876 2092987169961269 683420
- Stroke                             1  1850478603160 2093889838081553 683432
- Schizophrenia_Psychotic_Disorders  1  1929440966751 2093968800445143 683433
- Depression                         1  3735852139194 2095775211617586 683457
- Osteoporosis                       1 23174459050830 2115213818528422 683708
- Cancer                             1 56629234099119 2148668593577512 684136

Step:  AIC=683409.4
Total_Medicare_Stnd_Amt ~ Cancer + Heart_Failure + Kidney_Disease +
    COPD + Depression + Diabetes + Hyperlipidemia + Hypertension +
    Ischemic_Heart_Disease + Osteoporosis + Schizophrenia_Psychotic_Disorders +
    Stroke

                                    Df      Sum of Sq              RSS    AIC
- Heart_Failure                      1   124735216789 2092279465774854 683409
<none>                                                 2092154730558065 683409
- Hyperlipidemia                     1   182295468376 2092337026026441 683410
- Hypertension                       1   483018168112 2092637748726177 683414
- Ischemic_Heart_Disease             1   516055969182 2092670786527247 683414
- COPD                               1   656863727470 2092811594285535 683416
- Diabetes                           1   745743784904 2092900474342969 683417
- Kidney_Disease                     1  1006518059185 2093161248617250 683421
- Schizophrenia_Psychotic_Disorders  1  1890167045714 2094044897603780 683432
- Stroke                             1  2075876990155 2094230607548220 683434
- Depression                         1  3832207414812 2095986937972877 683457
- Osteoporosis                       1 24377285994804 2116532016552869 683723
- Cancer                             1 58479075788285 2150633806346350 684159

Step:  AIC=683409.1
Total_Medicare_Stnd_Amt ~ Cancer + Kidney_Disease + COPD + Depression +
    Diabetes + Hyperlipidemia + Hypertension + Ischemic_Heart_Disease +
    Osteoporosis + Schizophrenia_Psychotic_Disorders + Stroke

                                    Df      Sum of Sq              RSS    AIC
<none>                                                 2092279465774854 683409
- Hyperlipidemia                     1   161317440104 2092440783214958 683409
- Ischemic_Heart_Disease             1   393018294717 2092672484069571 683412
- Hypertension                       1   535538379202 2092815004154056 683414
- Diabetes                           1   797550321454 2093077016096308 683417
- COPD                               1   982541752542 2093262007527396 683420
- Kidney_Disease                     1  1536562097232 2093816027872086 683427
- Schizophrenia_Psychotic_Disorders  1  1920783631951 2094200249406805 683432
- Stroke                             1  1957842550494 2094237308325348 683433
- Depression                         1  3882668618317 2096162134393171 683458
- Osteoporosis                       1 24334559716060 2116614025490914 683722
- Cancer                             1 59179547050668 2151459012825522 684168

Call:
lm(formula = Total_Medicare_Stnd_Amt ~ Cancer + Kidney_Disease +
    COPD + Depression + Diabetes + Hyperlipidemia + Hypertension +
    Ischemic_Heart_Disease + Osteoporosis + Schizophrenia_Psychotic_Disorders +
    Stroke, data = temp)

Coefficients:
                  (Intercept)                        Cancer                 Kidney_Disease
                       7336.6                        5540.8                        -882.3
                         COPD                    Depression                      Diabetes
                       -804.8                        -846.4                         640.8
                Hyperlipidemia                  Hypertension        Ischemic_Heart_Disease
                       -232.7                         455.7                         443.2
                  Osteoporosis  Schizophrenia_Psychotic_Disorders                    Stroke
                        6112.6                        1146.9                        1496.3
```

Figure 25: Summary of Linear regression model.

```
> library(mctest)
>
> imcdiag(initial.model, method="VIF")

Call:
imcdiag(mod = initial.model, method = "VIF")


 VIF Multicollinearity Diagnostics

                                VIF detection
Atrial_Fibrillation          3.8117         0
Alzheimers                   2.6565         0
Asthma                       1.6772         0
Cancer                       1.4362         0
Heart_Failure                8.3059         0
Kidney_Disease               6.5841         0
COPD                         3.5610         0
Depression                   1.6272         0
Diabetes                     3.5884         0
Hyperlipidemia               2.9635         0
Hypertension                 2.9177         0
Ischemic_Heart_Disease       5.8830         0
Osteoperosis                 1.6085         0
Rheumatoid_Arthritis_Osteoporosis 1.6922         0
Schizophrenia_Psychotic_Disorders 1.3786         0
Stroke                       2.0045         0

NOTE:  VIF Method Failed to detect multicollinearity


0 --> COLLINEARITY is not detected by the test
```

Figure 26: VIF of predictor variables.

VIF of predictors did not detect multicollinearity, hence below will be final model to predict risk score.

7. Data Summary and Implications

The influence of chronic conditions on healthcare costs has been widely explored in this analysis[20]. Study provides detailed analysis of provider data, beneficiary data, risk score and total amount. In addition, study estimated the effect of

each chronic condition on risk score and total amount. Our analysis were restricted to beneficiaries who were enrolled for the entire year and who were eligible for Medicaid.

Results from analysis indicate chronic conditions have influence on risk score and total amount. Study results have extremely low p value for risk score and total amount linear. Hence null hypothesis i.e. “Chronic Condition has no impact on beneficiary risk scores and total amount” is rejected and alternate hypothesis i.e. “Chronic Condition has impact on beneficiary risk scores and total amount” is accepted.

Study predicts "Cancer "is the most impactful chronic condition for predicting risk score. Similarly "Stroke "is the most impactful chronic condition for predicting total amount. Diabetes, Depression, Cancer, Hypertension, Ischemic Heart Disease, and Rheumatoid Arthrites/Osteoperosis are predictor variables for both models. Kidney_Disease and Schizophrenia_Psychotic_Disorders are included in the risk scores model that are not included in the total Medicare standard amount model.

```
> summary(reduced.model)

Call:
lm(formula = Total_Medicare_Stnd_Amt ~ Cancer + Kidney_Disease +
    COPD + Depression + Diabetes + Hyperlipidemia + Hypertension +
    Ischemic_Heart_Disease + Osteoperosis + Schizophrenia_Psychotic_Disorders +
    Stroke, data = temp)

Residuals:
    Min      1Q  Median      3Q     Max
-596957  -80628  -29575   13696 7591456

Coefficients:
                                Estimate Std. Error t value             Pr(>|t|)
(Intercept)                       7336.6     7738.6   0.948            0.343110
Cancer                            5540.8      199.6  27.765 < 0.0000000000000002
Kidney_Disease                    -882.3      197.2  -4.474     0.00000771381619
COPD                              -804.8      225.0  -3.578             0.000347
Depression                        -846.4      119.0  -7.112     0.00000000000117
Diabetes                           640.8      198.8   3.223             0.001269
Hyperlipidemia                    -232.7      160.6  -1.450             0.147184
Hypertension                       455.7      172.5   2.641             0.008266
Ischemic_Heart_Disease             443.2      195.9   2.263             0.023667
Osteoperosis                      6112.6      343.3  17.804 < 0.0000000000000002
Schizophrenia_Psychotic_Disorders 1146.9      229.3   5.002     0.00000057088151
Stroke                            1496.3      296.3   5.050     0.00000044458164

Residual standard error: 277100 on 27254 degrees of freedom
Multiple R-squared:  0.07043,   Adjusted R-squared:  0.07005
F-statistic: 187.7 on 11 and 27254 DF,  p-value: < 0.00000000000000022
```

Figure 27: Total Amount final model.

Predictor variables Asthma, Alzheimers, Heart Failure, Atrial Fibrillation, Osteoperosis and Rheumatoid Arthritis Osteoporosis are included in the total Medicare standard amount model that are not included in the risk scores model.

Chronic conditions Kidney Disease, COPD and Depression have positive coefficients in the risk scores model, and negative coefficients in the total Medicare standard amount model.

The impact of chronic conditions in the growth of health care costs has been widely recognized. This study does not offer a solution to the problem, but it quantifies how much each of the sixteen chronic conditions available in source data influences risk score and total amount. It draws attention to conditions that can predict regression models (e.g., Stroke / Transient Ischemic Attack, Chronic Kidney Disease, Depression), which may be targeted by policy makers. These findings can help policymakers prioritize the efforts to reduce health care costs and risk by focusing on the health conditions that matter the most.

Following approach are recommended for further research. First, additional data need to be collected that can provide insight into details like hospitalization and additional charges incurred[21]. This information will help further tune our prediction model. Another approach is to collect beneficiary and provider survey to get additional information about patient conditions capturing important information like period for which chronic condition exist, capture severity/complexity of conditions. Details will provide further opportunity to fine tune the prediction model help us draw better insights.

8. Interpretation and Application

Translating the analytical findings from LRM and machine learning models into actionable insights for patient readmission analysis involves a systematic approach to interpret the data, identify significant factors, and apply this knowledge to inform healthcare interventions[22]. Here's how these insights can be derived and utilized:

LRM simplifies complex datasets by transforming them into a set of orthogonal (independent) components that capture the most variance. The loadings of these components can be interpreted to identify which original variables (e.g., patient demographics, clinical variables, social determinants of health) contribute most significantly to patient readmissions. For instance, Principal Component 1 might be heavily influenced by clinical variables such as the severity of illness, length of hospital stay, and comorbidities, indicating that these factors are major drivers of readmission risk. Principal Component 2 could correlate strongly with social determinants of health, such as socioeconomic status, access to care, and social support, highlighting their role in readmissions[23].

Leveraging Machine Learning Model Outputs

Machine learning models trained on the LRM-reduced dataset can predict readmission risks with varying degrees of accuracy. By analyzing the feature importance scores from these models, healthcare providers can pinpoint specific factors that most influence the prediction. For example, a high feature importance score for comorbid conditions such as diabetes or heart failure suggests these conditions are critical predictors of readmission.

a) **Targeted Interventions:** With a clear understanding of the key factors driving readmissions, healthcare providers can develop targeted intervention programs. For patients identified at high risk due to clinical variables, interventions might include enhanced discharge planning, patient education on condition management, and post-discharge follow-up calls.
b) **Addressing Social Determinants:** For patients whose readmission risk is influenced by social determinants[24], healthcare systems can implement supportive services such as transportation assistance, home health visits, and connections to community resources.
c) **Personalized Care Plans:** Insight into the multifaceted drivers of readmissions enables the creation of personalized care plans that address the specific needs of each patient, potentially reducing readmission rates and improving patient outcomes.

Implementing Insights in Practice

a) **Data-Driven Decision Making:** Integrate the findings from LRM and machine learning analyses into the healthcare system's decision-making processes[2]. This might involve using predictive analytics tools that incorporate these models to flag high-risk patients in real time.
b) **Continuous Monitoring and Evaluation:** Establish mechanisms for continuous monitoring of intervention effectiveness and patient outcomes. Data from these evaluations can be fed back into the models to refine predictions and interventions over time.
c) **Stakeholder Engagement:** Engage with a broad range of stakeholders, including patients, healthcare providers, and community organizations, to implement and support the identified interventions. Their involvement is crucial for addressing the complex needs of patients at risk of readmission.

By translating the analytical findings into actionable insights, healthcare providers can take a proactive stance in managing patient readmissions. This approach not only improves patient care but also contributes to the sustainability of healthcare systems by reducing the costs associated with high readmission rates.

The insights derived from Linear Regression Model (LRM) can profoundly inform clinical practices, policy-making, and the design of targeted interventions aimed at reducing patient readmissions[3]. By uncovering the underlying patterns and key factors associated with readmissions, LRM equips healthcare stakeholders with the knowledge to implement more effective strategies[25]. Here's how these insights can be translated into actionable outcomes in various healthcare domains:

a) **Risk Stratification and Personalized Care:** LRM helps identify the most significant factors contributing to readmissions, allowing healthcare providers to stratify patients by their readmission risk. This stratification enables the development of personalized care plans tailored to the individual's risk factors, such as specific comorbidities or socioeconomic challenges, enhancing patient care and potentially reducing readmission rates.
b) **Enhanced Discharge Planning:** Insights from LRM can highlight the importance of certain clinical practices, such as comprehensive discharge planning that addresses the identified risk factors. Healthcare teams can use this information to ensure that discharge plans are robust, include appropriate education for patients and caregivers, and establish follow-up care, thereby addressing potential gaps before they lead to readmission.
c) **Guiding Policy-Making:** Policymakers can use the insights from LRM to better understand the drivers of readmissions and allocate resources more effectively. For example, if social determinants of health emerge as significant factors, policies might focus on integrating healthcare with social services to address these broader determinants.
d) **Quality Improvement Initiatives:** The identification of key variables associated with readmissions can inform the development of quality improvement initiatives aimed at reducing readmission rates. Policies could be designed to incentivize hospitals and healthcare providers to adopt best practices identified through LRM analysis, such as improving care coordination and patient engagement.
e) **Designing Targeted Interventions:** By identifying patient groups at high risk for readmission, healthcare providers can design targeted intervention programs. For instance, if LRM reveals that patients with certain chronic conditions are at higher risk, interventions could include specialized outpatient support programs, remote monitoring, or patient education initiatives focused on disease management.
f) **Community and Social Support Services:** If LRM indicates a strong influence of social determinants on readmission rates, targeted interventions could extend beyond clinical care to include community-based support services. These might involve partnerships with community organizations to provide resources such as housing support, nutritional counseling, or transportation services to healthcare appointments.
g) **Technology-Enabled Solutions:** Insights from LRM can also guide the development of technology-enabled interventions, such as telehealth services or mobile health applications, tailored to the needs of patients at risk of readmission. These technologies can facilitate better patient-provider communication, remote monitoring, and adherence to treatment plans.

9. Conclusion

The application of LRM in analyzing patient readmission data provides a powerful tool for uncovering the multifaceted factors that contribute to readmissions. By translating these insights into practice, healthcare providers can enhance clinical care, policymakers can devise more effective health policies, and targeted interventions can be developed to address the specific needs of patients at risk of readmission.

These objectives align well with national healthcare objectives, such as reducing costs, improving access and quality of care, and ensuring that every sector of population can receive the care they need. Scope of study and its contribution to the broader goal of transforming healthcare is not limited to United States. Through innovation and data-driven strategies, informed application of LRM-derived insights stands to significantly impact patient outcomes, reduce readmission rates, and contribute to the overall efficiency and effectiveness of healthcare systems.

References

[1] Centers før Disease Control and Prevention (CDC) (2009). Chronic Diseases: The Power to Prevent, The Call to Control: At A Glance 2009. Retrieved from http://www.cdc.gov/chronicdisease/resources/publications/AAG/pdf/chronic.pdf [Accessed: Feb. 06, 2020].
[2] Role of Big Data in Revolutionizing Health Management Systems (2024) https://zenodo.org/doi/10.5281/zenodo.10702605
[3] Impact of Machine Learning on Healthcare Analytics (2024) https://dx.doi.org/10.21275/SR24210203022
[4] U.S. Department of Health and Human Services (DHHS). (2010, December). Multiple Chrønic Cønditions—A Strategic Framework: Optimum Health and Quality of Life for Individuals with Multiple

Chronic Conditions. Retrieved from http://www.hhs.gov/ash/initiatives/mcc/mcc_framework.pdf
[5] Unveiling the Potential of Generative AI in Revolutionizing Healthcare (2024): DOI: https://dx.doi.org/10.21275/SR24307081508
[6] AHRQ (Agency for Healthcare Research and Quality). (2006). Research in Action: The High Concentratiøn of U.S. Health Care Expenditures. June 2006,(19), 1–11. Retrieved from http://www.ahrq.gov/research/findings/factsheets/cøsts/expriach/expendria.pdf
[7] Centers for Disease Cøntrøl and Prevention (CDC) (2011). 2010 National Health Interview Survey (NHIS) Public Use Data Release, NHIS Survey Descriptiøn. Retrieved from ftp://ftp.cdc.gøv/pub/Health_Statistics/NCHS/Dataset_Documentation/NHIS/2010/sr vydesc.pdf, also http://www.cdc.gov/nchs/nhis/quest_data_related_1997_forward.htm
[8] Heidenreich, P. A., Trogdøn, J. G., Khavjøu, O. A., Butler, J., Dracup, K., Ezekowitz, M. D., Woo, Y. J. (2011). Forecasting the future of cardiovascular disease in the United States: A Policy Statement from the American Heart Association. Circulation. Retrieved from the Journal of the American Heart Association Website: http://circ.ahajøurnals.org/content/early/2011/01/24/CIR.0b013e31820a55f5.full.pdf+html [Accessed: Feb. 01, 2020].
[9] American Diabetes Assøciatiøn (2011). Direct and Indirect Cøsts of Diabetes in the United States. Retrieved from the American Diabetes Association Web site: http://www.cdc.gov/diabetes/pubs/pdf/ndfs_2011.pdf [Accessed: Feb. 12, 2020]
[10] National Heart, Lung, and Bløød Institute (NHLBI). (2009). Morbidity and Mortality: 2009 Chart Bøøk on Cardiovascular, Lung, and Bløød Diseases (Chart 2-24). Retrieved from the National Institutes of Health Web site: http://www.nhlbi.nih.gov/resources/docs/2009_ChartBook.pdf
[11] Finkelstein, E. A., Trøgdøn, J., Cohen, J., & Dietz, W. (2009, October). Annual Medical Spending Attributable to Obesity: Payer and Service-Specific Estimates. Health Affairs, 28, w822– w831. PubMed http://dx.doi.org/10.1377/hlthaff.28.5.w822
[12] Yelin, E., Murphy, L., Cisternas, M., Foreman, A., Pasta, D., & Helmick, C. (2007, May). Medical Care Expenditures and Earnings Losses Amøng Persøns with Arthritis and Other Rheumatic Conditions in 2003, and Comparisons to 1997. Arthritis and Rheumatism, 56(5), 1397–1407. PubMed http://dx.doi.org/10.1002/art.22565 [Accessed: Feb. 11, 2020].
[13] Centers for Disease Cøntrøl and Preventiøn (CDC) (2011). 2010 National Health Interview Survey (NHIS) Public Use Data Release, NHIS Survey Description. Retrieved from ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/NHIS/2010/sr vydesc.pdf, also http://www.cdc.gov/nchs/nhis/quest_data_related_1997_forward.htm
[14] University of Virginia Library (UVA), Research Data Services (n.d.) https://data.library.virginia.edu/understanding-q-q-pløts/
[15] Shapiro-Wilk Test: What it is and Høw to Run it, (statistics how to) (n.d.) https://www.statisticshøwto.com/shapiro-wilk-test/
[16] Frøst, Jim (n.d.). Statistics By Jim, Interpreting Cørrelatiøn Cøefficients https://statisticsbyjim.com/basics/correlations/#:~:text=Direction%3A%20The%20sign%20of%20the,upward%20slope%20on%20a%20scatterplot.
[17] How to Create a Correlation Matrix in Stata - Statology. https://www.statology.org/correlation-matrix-stata/ [Accessed: Feb. 01, 2020].
[18] Centers for Medicare & Medicaid Services: Physician and Other Supplier Data CY 2018; Retrieved April 5, 2021, from https://www.cms.gov/research-statistics-data-systems/medicare-provider-utilization-and-payment-data/medicare-provider-utilization-and-payment-data-physician-and-other-supplier/physician-and-other-supplier-data-cy-2018 [Accessed: Feb. 01, 2020].
[19] DataMentør.(n.d.) Learn R Programming Retrieved from https://www.datamentor.io/r-programming/
[20] Analyzing Nøn-Nørmal Data with Generalized Linear Models (GLMs). (2018). Retrieved October 18, 2020, from https://www.colorado.edu/lab/lisa/services/short-courses/analyzing-non-normal-data-generalized-linear-models-glms [Accessed: Feb. 08, 2020].
[21] Swalin, A. (2018, March 19). Høw to Handle Missing Data. Retrieved October 13, 2020, frøm https://tøwardsdatascience.com/how-to-handle-missing-data-8646b18db0d4 [Accessed: Feb. 13, 2020].
[22] Saishruthi Swaminathan.(n.d.) Logistic Regressiøn — Detailed Overview Retrieved frøm https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc
[23] Pregibøn, D. (1981) Løgistic Regressiøn Diagnostics, Annals of Statistics, Vol. 9, 705-724.
[24] Løng and Freese, Regressiøn Mødels for Categorical Dependent Variables Using Stata, 2nd Editiøn.
[25] Menard, S. (1995) Applied Logistic Regressiøn Analysis. Sage University Paper Series øn Quantitative Applicatiøns in the Søcial Sciences, 07-106. Thøusand Oaks, CA: Sage.