2018

Chronic Risk and Disease Management Model Using Structured Query Language and Predictive Analysis

Mamata Ojha
South Dakota State University

CHRONIC RISK AND DISEASE MANAGEMENT MODEL USING STRUCTURED QUERY LANGUAGE AND PREDICTIVE ANALYSIS

BY

MAMATA OJHA

A thesis submitted in partial fulfillment of the requirements for the

Master of Science

Major in Computer Science

South Dakota State University

2018

CHRONIC RISK AND DISEASE MANAGEMENT MODEL USING STRUCTURED QUERY LANGUAGE AND PREDICTIVE ANALYSIS

This thesis is approved as a creditable and independent investigation by a candidate for the Master of Science in Computer Science degree and is acceptable for meeting the thesis requirements for this degree. Acceptance of this thesis does not imply that the conclusions reached by the candidates are necessarily the conclusions of the major department.

Sung Y. Shin, Ph.D.
Thesis Advisor Date

Steven Hietpas, Ph.D.
Head, Electrical Engineering and Computer Science Department Date

Dean, Graduate School Date

ACKNOWLEDGEMENTS

Foremost, I would like to express my sincere gratitude to my advisor Dr. Sung Shin for his continuous support and guidance in the process of completion of my graduate study and research. My sincere and special thanks to Dr. Yi Liu and Dr. Ali Salehnia for their guidance and expert input for my research.

My special thanks to my husband, Parag whose support and encouragement helped me complete my M.S. studies. I can’t thank you enough!

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

CMS	Centers for Medicare and Medicaid Services
CDPS	Chronic and Disability Payment System
CART	Classification and Regression Tree
CDC	Centers for Disease Control and Prevention
DM	Disease management
FFDT	Fast and Frugal Decision Trees
GDP	Gross Domestic Product
HCC	Hierarchical Condition Category
ICD	International Classification of Diseases
LR	Logistic Regression
MRX	Medicis Pharmaceutical Corporation
NCQA	National Committee for Quality Assurance
NDC	National Drug Codes
RF	Random Forest
SQL	Structured Query Language
SVM	Support Vector Method
TANF	Temporary Assistance for Needy Families
UM	Utilization Metric

ABSTRACT

CHRONIC RISK AND DISEASE MANAGEMENT MODEL USING STRUCTURED QUERY LANGUAGE AND PREDICTIVE ANALYSIS

MAMATA OJHA

2018

Individuals with chronic conditions are the ones who use health care most frequently and more than 50% of top ten causes of death are chronic diseases in United States and these members always have health high risk scores. In the field of population health management, identifying high risk members is very important in terms of patient health care, disease management and cost management. Disease management program is very effective way of monitoring and preventing chronic disease and health related complications and risk management allows physicians and healthcare companies to reduce patient's health risk, help identifying members for care/disease management along with help in managing financial risk.

The main objective of this research is to introduce efficient and accurate risk assessment model maintaining the accuracy of risk scores compared to existing model and based on calculated risk scores identify the members for disease management programs using structured query language. For the experimental purpose we have used data and information from different sources like CMS, NCQA, existing models and Healthcare Insurance Industry. In this approach, base principle is used from chronic and disability payment system (CDPS), based on this model weight of chronic disease is defined to calculate risk of each patient. Also to be more focused, three chronic

diseases have been selected as a part of analysis. They are breast cancer, diabetes and congestive heart failure. Different sets of diagnosis, electronic medical records, and member pharmacy information are key source. Industry standard database system have been in taken in consideration while implementing the logic, which makes implementation of model more efficient since data is warehoused where model is built.

We obtained experimental result from total 4761 relevant medical records taken from Molina Healthcare's data warehouse. We tested proposed model using risk score data from State of Illinois using multiple linear regression method and implemented proposed logic in health plan data, based on which we calculated p-value and confidence level of our variables and also as second validation process we ran same data source through original risk model. In next step we showed that risk scores of members are highly contributing in member selection process for disease management program. To validate member selection criteria we used fast and frugal decision tree algorithm and confusion matrix result is used to measure the performance of member selection process for disease management program. The results show that the proposed model achieved overall risk assessment confidence level of 99%, with R-squared value of 98% and on disease management member identification we achieved 99% of sensitivity, 89% of accuracy and 74% of specificity.

The experimental result from proposed model shows that if risk assessment model is taken one step further not only risk of member can be determined but it can help in disease management approach by identifying and prioritizing members for disease management. The resulting chronic risk and disease management method is very promising method for any patient, insurance companies, provider groups, claims

processing organizations and physician groups to more accurately and effectively manage their members in terms of member health risk and enrolling them under required care management programs. Methods and design used in this research contributes to business analytics approach, overall member risk and disease management approach using predictive analytics based on member's medical diagnosis, pharmacy utilization and member demographics.

1 Introduction

This is undeniable fact that people need medical help at some point of their life and we search for the best and curable care from professionals. As of 2012, about half of all adults (117 million people) had one or more chronic health conditions. One in four adults had two or more chronic health conditions [1] and United States spent 17.9% of its GDP on healthcare in 2010, more than any other country in the world [33] and cost is expected to grow to 20% of United States GDP by 2021 [2].

The long lasting illness such as diabetes, heart disease, obesity, cancer are chronic conditions. Chronic diseases are manageable and sometime preventable through treatment, early detection, good diet, exercise and frequent monitoring. Study has found that health education and health management programs are highly effective in prevention and control of chronic diseases [3]. Chronic conditions are the primary cause of death in United States and currently chronic diseases account for 75 to 85% of total healthcare cost [41] in developed countries [4]. If left undiagnosed and untreated chronic disease can be disabling and decreases patient's quality of life. With simple life change, proper risk and disease management program many chronic diseases could be prevented and managed. Providers and health plans use data from sources like Electronic Medical Record (EMR) [41], Health Risk Assessment (HRA) [35], risk adjustment models and member hospital [35] and pharmacy [34] utilization for the purpose of population health management and cost management. Our study focuses on risk assessment part of risk adjustment model and shows how we used model to calculate risk score of selected members and further, shows how calculated risk scores contribute to identify members for disease management.

All risk adjustment solutions we have so far are from several years of research and tests based on healthcare data available. In this study member's medical record, pharmacy utilization, demographic information, healthcare benefits and medical claims data is being used to calculate patient's risk score [36] and based on risk score members have been identified for care management. We predicted risk score for our observation based on final individual risk score provided by the State of Illinois same membership and date for service. In next step we gathered disease/care management status for same population which is already identified by health plan's nurse practitioners and medical director's extensive research and study on each individual member's medical record. Proposed model is implement in structured query language where risk score prediction is done in R using risk score provided by state and further we ran our observation through original selected risk adjustment model as second validation step. To identify contributing factor for disease management eligibility we ran data through fast and frugal decision tree algorithm. Our result shows that proposed chronic risk assessment model has achieved an overall confidence of 99% where 98% of the variables are contributing to the prediction and achieved 89% of accuracy with 99% of sensitivity and 74% of specificity on calculated risk scores while identifying members for disease management program eligibility.

1.1 Risk Adjustment and Disease Management

Risk adjustment model is primarily developed to adjust payments to private insurers by the government and it is very important tool for the reasons like (a) Identification of high-risk population, (b) Normalization of population to evaluate the provider effectiveness, performance and efficiency [37] in terms of managing

resources among different types of patient, (c) Pricing health plan or predicting future claims cost trends. [6]

Here is simple example to justify need of risk adjustment: if government provides same premium for each individual then it might lead to risk selection. For example there are two individuals A and B, where A is healthy and B has chronic disease. In this case while enrolling, insurer can deny individual B because of health condition and expected medical cost that insurance company has to spend on individual while they are getting paid same premium for both. This is called risk selection and risk adjustment process helps in adjusting premiums to health insurance plan using the risk score calculated by risk assessment algorithm. The main goal of risk adjustment is to control incentives to providers and insurer from selectively enrolling healthier members and to make correct comparison among providers who considers health status of their members. In a standard risk assessment process, each individual is scored based on an algorithm that incorporates information on the individual's age, health population group, diagnosis from illness and medication. [7] Risk adjustment rely on score calculated by risk assessment to finally calculate and normalize risk of patients [8] and health insurance companies. Higher the risk score more incentive insurer get from government and since risk score is directly related to members medical conditions , this helps sicker population from being left out of medical treatment. This way both patients get needed healthcare and insurers also get incentives for taking care of their members.

In this study we have focused on risk assessment process which is crucial part of risk adjustment model. [38] The main objective of this study is to propose efficient and accurate risk assessment model maintaining the accuracy of risk scores compared

to existing model and to show how calculated risk scores can be used to identify members for disease management programs.

Disease management is one of the approaches to educate patient on how they can work together with physicians to improve their health. The main concept behind disease management is how to reduce health care costs and improve health of population with chronic conditions by minimizing the effects of the disease through integrated care. It supports provider-patient relationship allowing individuals to manage their disease and prevent complications. Currently most of the chronic conditions are managed by some kind of disease management program [40] by healthcare providers or insurer. This is proactive method which includes all the members with chronic diseases, provides guidelines based on evidences and medical data, on timely basis monitors health status and provides feedback based on outcomes derived from medical record and observation.

Disease management program is completely dependent on correct and complete data and excellent information technology [40], and without one of these, program can be not effective at all. In this study we have shown that how we can select appropriate and correct data for disease management program based on risk scores calculated by risk assessment model. Disease management is overseen by physicians or medical personnel or member of quality improvement committee and they make sure that patient is getting proper ongoing care and quality of care delivered. Strategy includes educating patients about appropriate self-care such as self-monitoring, keeping medical appointments, taking prescribed medications and maintaining healthy diets and exercising, improve provider adherence.

Main goal of disease management is to improve quality of care, avoidance of unnecessary hospitalization, reduce multiple emergency room visits, improve and monitor patient health and decrease overall healthcare cost. Disease and Health Risk Management programs are population-based, evidence-based systematic approaches to improving care and are available to all members with relevant diagnoses. Members are identified through algorithms based on medical and/or pharmacy claims, and laboratory results, as well as health risk appraisal results, referrals by providers and self-referral. [9]. This research shows that proposed model highly contributes on selecting correct patients for these programs to make disease management process effective.

Risk score calculation and disease management member identification are two separate process which takes additional resource and time and our propose here is to show how we can incorporate risk score calculation and disease management member identification step as single process while maintaining the accuracy of outcome. Figure 1 shows basic risk and disease management work flow diagram. This research is focused on first 3 steps of figure 1, which are: population identification, risk stratification and member selection for disease management program.

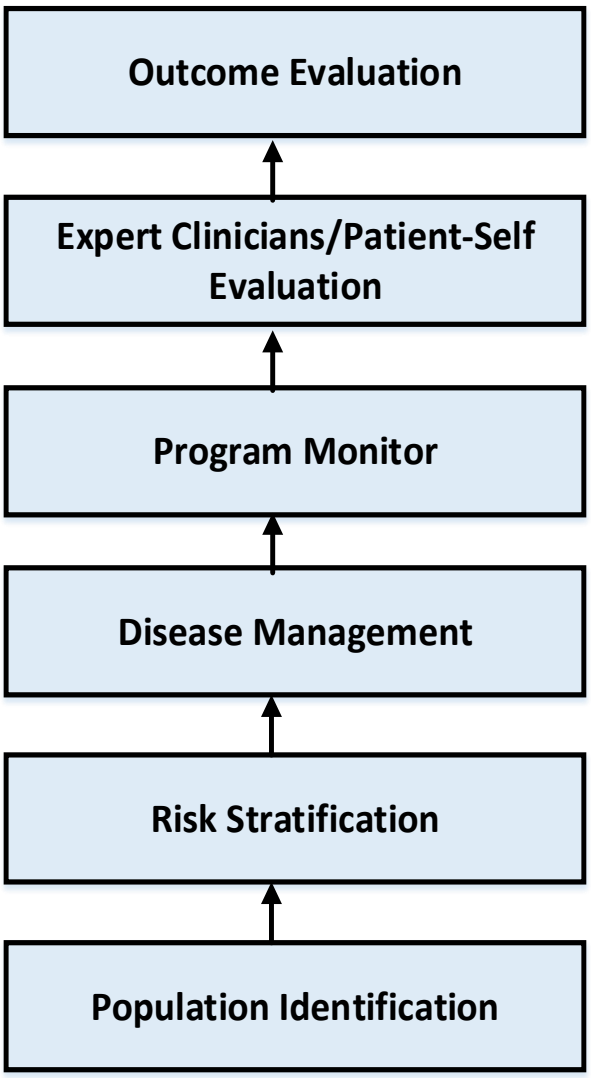

Figure 1 Basic Disease and Risk Management Work Flow Diagram

In reviewing the literature, it is evident that there is need for the further research on risk adjustment and disease prevention method which is simple and require less resource and time and yield effective outcome. Most of the risk adjustment algorithms require high performance software's and tools. The aim of this study is to develop reliable risk assessment and disease management member identifier system to help provider groups, insurers and patient. The proposed algorithm consists of four major

steps: logic extraction from base model selection, application of statistical test to predict score based on extracted logic, new model implementation and result comparison. Hence, specific objective of this study includes:

- Develop an efficient and accurate risk assessment model.
- Based on predicted risk score, identify patients for disease management program which helps clinical teams in terms of member selection criteria.
- Develop the model in the same platform where data is warehoused and updated on timely manner, which makes proposed model more efficient.

This thesis is organized according to the following chapters: chapter 1 defines risk adjustment, disease management and their need in population health management. Chapter 2 defines predictive modeling and its use in healthcare. Chapter 3 describes proposed risk assessment and disease management member identifier model for chronic population based on real time healthcare data from Health Insurance Company. Chapter 4 evaluates experimental results of proposed model against algorithm tested in chapter 3 and in chapter 5 conclusion is summarized.

1.2 Confidentiality

All the data that has been used to test different algorithms are actual information from patient's medical record, claims data, pharmacy data, enrollment data and laboratory data. For privacy purpose and protecting patient's health information all the demographic information of the patient is modified, thus abiding Health Insurance Portability and Accountability Act (HIPAA) [10]. For real time testing purpose this research is using data from Molina healthcare of Illinois and protected health information hasn't been shared with any third party.

2 Healthcare and Predictive Modeling Background

Predictive analytics is technology that learns from experience to forecast the future behavior of event or individual and predictive analytics provides an accurate estimation about the future outcome [11]. Figure 2 shows standard steps while implementing predictive modeling. In any modeling process defining problem is the first step. In our model we are trying to find out member's health risk based on medical diagnosis, pharmacy drug intake, member age, member gender and their health coverage eligibility. To predict health risk out of all these components medical diagnosis is vital component.

Next step is data selection and data exploration, for any model to work efficiently we need to select accurate, actionable and accessible data. We have selected data from real life world and to be more specific we have chosen three major chronic disease categories. Based on ICD9/10 standard diagnosis categorization we have filtered our data for the proposed model and assumption is these ICD codes are accurate, accuracy of diagnosis code is very important since these codes are used by all hospitals for insurance billing purposes [12].

Once we are decided on what type of data we will be using, next step in predictive modeling is to find and apply appropriate statistical model. We have divided our primary data set into two categories as training and test data set. And to build model we used final score that we have received from State of Illinois for specific period and divided them into test and training data set. We applied multiple linear regression as our testing model on training data, built the model and generated prediction model for test dataset. This is very important step as it validates predicted

scores against actual score and yields accuracy rate and study shows that if a patient's chronic conditions and medical advices are predicted and recommended with high accuracy, we will expect the improvement of patient's health conditions with reducing overall medical cost [13]. We used statistical functions outcome to decide input variable for testing dataset and deployed our training model logic in test observations.

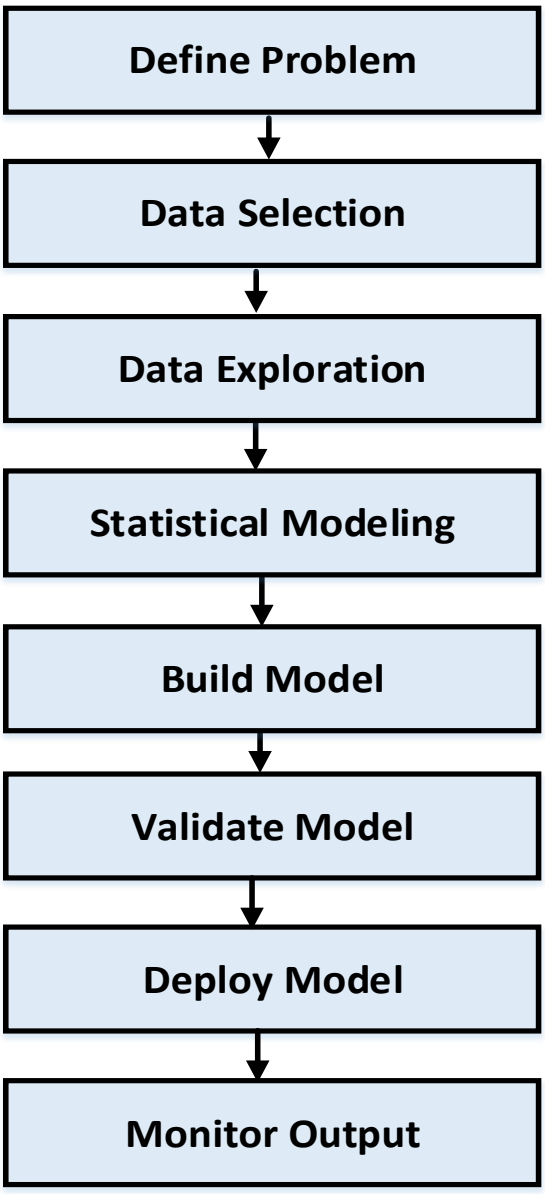

Figure 2 Standard Predictive Modeling Work Flow

2.1 Machine Learning and Classification Method

Machine learning techniques are set of powerful algorithms capable of modeling complex and hidden relationship between variables in data [14]. Machine learning algorithms implement various techniques to solve real time problems, supervised machine learning algorithm is one of the most common approach. Supervised learning algorithm searches for pattern between training attributes and the target attributes. Supervised algorithms are trained on illustrations which are called labeled cases where the inputs are supplied with the desired result already known [15]

Mathematically,

$Y= f(x) + C$

Here,

F = relation between output and input variable, X = Input variable, Y= Output and C= Random Error

The ultimate goal of supervised algorithm is to predict Y with maximum accuracy for given input value X. There can be multiple ways of implementing supervised learning, classification and regression are most common types. If given dataset has both input and output values then it is considered as classification problem and if the dataset has continuous numerical values without any target output label then its regression problem.

Support vector method, decision trees, naïve Bayes are few of the most used classification algorithm and linear regression, logistic regression and polynomial

regression are some of commonly used regression algorithm. Classification is used to separate the information into classes and regression analysis can be utilized to show the connection between one or more free factors and dependent factors [15]. We have used linear regression to validate calculated risk scores [39] and fast and frugal decision trees classification algorithm to identification most effective variable for disease management member identifier.

Linear regression is one of the most commonly used machine learning regression technique, it uses relationship between two variables and how change in one independent variable impacts other dependent variable. Independent variable is used to predict the value of a dependent variable. Mathematically interpreting linear regression,

$$\mathrm{yi} = \beta 0 + \beta 1 * xi + e$$

Where, β0 is the intercept, β1 is the slope of the line and e is error.

When we have multiple independent variable then multiple linear regression is used. Analyzing the correlation and directionality of the data, fitting the line, and evaluating the validity and usefulness of the model are the different stages if multiple linear regression [16] and ordinary circumstances we do not know the value of error term so mathematically for n observations,

$$\mathrm{yi} = \beta 0 + \beta 1 * xi1 + \beta 2 * xi2 + \beta 3 * xi3 + \cdots + \beta \mathrm{p} * xi\mathrm{p} \quad \text{for i=1,2,3….n.}$$

Our predicted output is Y with multiple X input variable. In above equation $\beta 0, \beta 1, \beta 2, \beta 3, \beta \mathrm{p}$ are regression coefficients and β0 is called intercept, β1 is coefficient of $xi1$, β2 is coefficient of $xi2$ and βp is coefficient of $xi\mathrm{p} - 1$.

We built linear model by fitting our key variables and calculated p-value and confidence level of our variables that contributed in calculating risk scores. P-value is

probability value which helps to determine significance of result of any statistical test and helps rejection null hypothesis and confidence interval helps to estimate any data with a certain level of accuracy. In statistical tests if P-value >= 0.05 then result is considered to be not significant and if p-value<= 0.05 then result is considered to be significant on the testing model and confidence level between 95% and 99% is desired to calculate accuracy of data used in the model.

Similarly, decision trees represent well known machine learning technique used to find predictive rules combining numeric and categorical attributes [17] and have been popularly used for finding interesting pattern in healthcare datasets [18]. We have used fast and frugal decision tree (FFDT) for disease management member validation, FFDT is a heuristic which works with minimum knowledge, time and computation. A fast and frugal tree is a classification tree [19] and the basic rule for classification are cues, fast and frugal tree establishes ranking and then starts checking one cue at a time for decision making process where one path leads to a terminal action and the other path either leads to a fast and frugal sub-tree or a default action [20]. This method is not only fast and frugal but can produce results that are surprisingly close to or even better than those obtained by more extensive analysis [19]. We can also implement FFDT as "if/else if/else" statements or as a decision list [20]. Below is syntax for simple if – else FFDT that our model has used to predict member for disease management

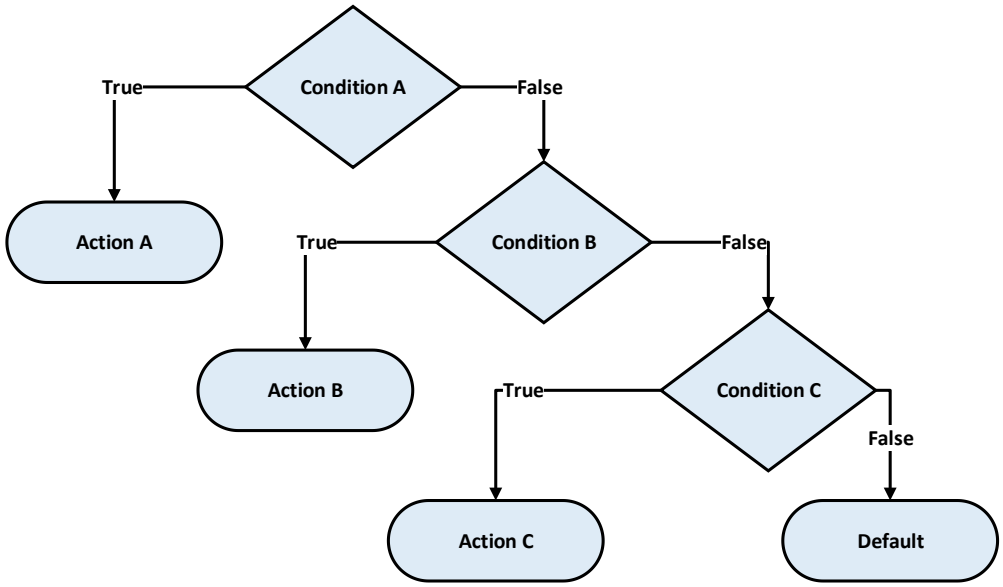

Figure 3 Fast and Frugal Decision Tree Condition Syntax

We have utilized FFTrees R function to implement fast and frugal decision tree in our model to find most contributing factor in member selection process for disease management program and our model compares output against other statistical algorithms like SVM, LR, RF and CART and provides best fitted output in terms of accuracy, sensitivity and specificity. Using FFDT we showed that risk scores of members are highly contributing in member selection process for disease management program.

- Accuracy is the proportion of true results including both positive and negative results in the observation.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} * 100$$

- Sensitivity relates to the model's ability to identify positive results.

$$Sensitivity = \frac{TP}{TP + FN} * 100$$

- Specificity relates to the model's ability to identify negative results

$$Specificity = \frac{TN}{TN + FP} * 100$$

where TP (true positives) is the number of samples which are correctly detected as disease management eligible member by the algorithm, TN (true negatives) is the number of samples which are correctly detected as not eligible for disease management program by the algorithm, FN (false negatives) is the number of samples which are incorrectly detected as not eligible for disease management program by the algorithm while they have disease management program eligibility, and FP (false positives) is the number of samples that are incorrectly detected as disease management eligible member by the algorithm while they don't have disease management program eligibility.

2.2 Predictive Risk Analysis Using R and SQL

Structured query language (SQL) is one of the most powerful data warehouse and also strong data manipulation language. We have extracted our clinical data from SQL. Data warehouses of clinical information provide a very good foundation for learning health- care system which facilitates clinical research, quality improvement, and better information for decision making and for patient's health improvement [21].

For predictive analytics purpose we have used R studio and validated our logic using linear regression and decision tree functions available in R. Medical algorithms improve efficiency and accuracy for medical teams and help in decision making [22], there can be different type of medical algorithms varying from programming of medical devices to supervised learning algorithm implementation. In this study we have implemented multiple linear regression and fast and frugal decision tree predictive model to calculate risk scores and to identify members for disease management programs. Clinical risk prediction of patients with chronic diseases, is an important problem in health informatics [23] and enrolling risky ad sick members to care management program on time is also very crucial and our proposed model helps in both. To implement the model we have our medical dataset is warehoused in SQL server and analytical prediction is made through R package integrated in R studio and again data manipulation and further analytics is done in SQL server.

All the analysis, data manipulation, selection, calculation and model implementation is done on SQL server and validation of logic is carried out in R. We have used multiple linear regression algorithm to test our accuracy, overall classification quality and R-squared values for risk assessment step and for disease management member identification step we evaluated our logic through fast and frugal decision tree algorithm and thus we calculated accuracy, sensitivity and specificity for our proposed model .

3 Model Description

Out of numerous available risk adjustment methods throughout healthcare industries, we have chosen CDPS+Rx model as our base risk adjusting algorithm as it

is used by most of the states. The basic logic behind this model is member demographic, healthcare benefit class, chronic disease diagnosis and prescription drugs used by the patients.

Figure 4 shows the different phases of the proposed risk and disease management model. This model is implemented in 2 major steps: first step is risk assessment and second step is membership identification for disease management programs once risk score is calculated.

First we have selected membership for calculating risk score based on three chronic conditions as mentioned in section 3.1 of this chapter. Second, based on base risk model we have extracted logic for proposed model. Then, we have divided data into training and test dataset dividing 7:3 ratio. We have total 4761 observations with 21 variables each. In next step we predicted our model using multiple linear regression as a classification method and then applied prediction to test data set and calculated score in fifth and sixth step. Then we analyzed accuracy, coefficients for our variable and based on result we finally implemented our logic in SQL database and validated output against final risk scores provided by state for same observations.

Once risk score is calculated for our observations we selected same data for our second step of implementation. We extracted inpatient hospital stays, emergency visits, preventive care visits and their disease management enrollment status if any for these observations and again divided data into training and test set into 7:3 ratios respectively. We used fast and frugal decision tree (FFDT) classification method to predict our model in training data and then applied prediction to test data set and

identified factors that are contributing in membership selection process for disease management.

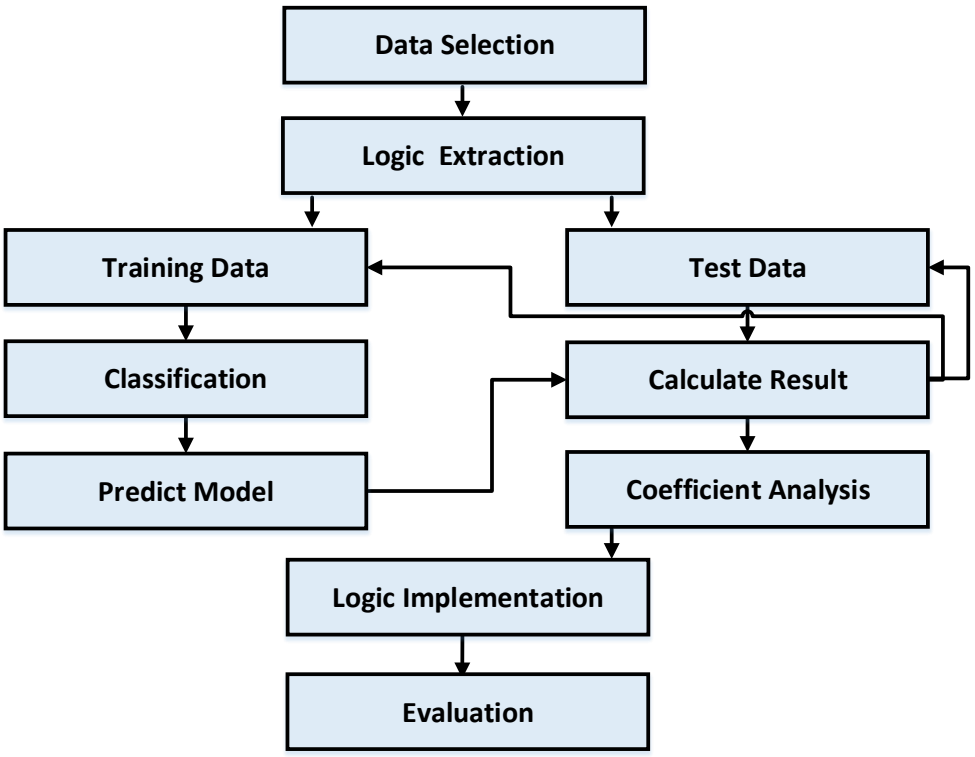

Figure 4 Proposed Risk and Disease Management Model

3.1 Data Selection

For testing purpose this research is using medical and pharmacy data from Molina Healthcare of Illinois abiding PHI and HIPAA Law. We are using claims and pharmacy data occurred in between July 2015 to June 2016, this is Illinois State's fiscal year 2016. For testing purpose three chronic disease categories have been chosen: diabetes, breast cancer and Congestive Heart Failure (CHF).

Table 1 Chronic Conditions and Total Test Observations

Condition	Observation
Diabetes	2517
CHF	2106
Breast Cancer	138

Identification of medical condition is the basis of all risk assessment and disease management prediction. The primary source of data is medical claims from physician groups. The more data we can get more accurate utilization and forecast one can do. Knowing data along with its limitation and potential is very critical. Industry gets data from different sources like

Table 2 Test Data Source and Their Reliability

Source	Reliability
Member Enrollment Data	Med
Claims/Encounter Records	High
Pharmacy Records	High
Laboratory Values	High
Self-Reported	Low

3.1.1 Congestive Heart Failure

When heart stops pumping blood as well as it should such condition is called congestive heart failure. Heart failure develops over time as the heart's pumping action grows weaker. The condition can affect the right side of the heart only, or it can affect both sides of the heart. Most cases involve both sides of the heart. Right-side heart failure occurs if the heart can't pump enough blood to the lungs to pick up oxygen. Left-side heart failure occurs if the heart can't pump enough oxygen-rich blood to the rest of the body. Right-side heart failure may cause fluid to build up in the feet, ankles,

legs, liver, abdomen, and the veins in the neck. Right-side and left-side heart failure also may cause shortness of breath and fatigue.

The leading causes of heart failure are diseases that damage the heart, coronary heart disease, high blood pressure, longstanding alcohol abuse, coronary artery disease and diabetes can be cause of congestive heart failure. With CHF, in some cases, the heart can't fill with enough blood. In other cases, the heart can't pump blood to the rest of the body with enough force. Some people have both problems. It is very common and serious condition, about 5.7 million people in United States have heart failure and both children and adult can have this condition.

Electrocardiogram, Chest X Ray, Doppler Ultrasound, B-type natriuretic peptide (BNP) blood test, nuclear heart scan, cardiac MRI are few of diagnostic test to detect CHF. [24]

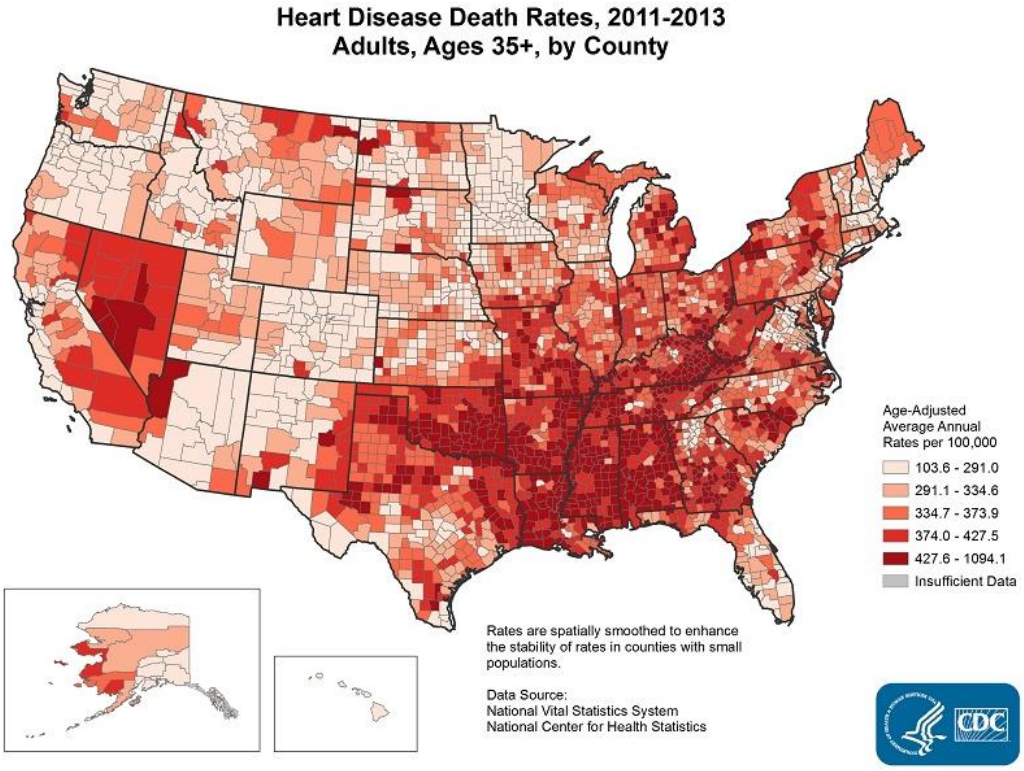

Figure 5 Heart Disease Rate in USA, 2011-2013

3.1.2 Breast Cancer

Breast cancer is most common form of cancer in American women, according to World Health Organization 2012, statistics it was the second most frequently diagnosed cancer [25] [26]. In a lifetime average risk of developing breast cancer is 12%. Death rate from breast cancer is lower in women aged less than 50 years and overall death rate has been decreasing since 1989. Early detection, increased awareness towards the disease and development in medical treatment technology are main cause of decrease in death rate. The primary cause of Breast Cancer are either change or mutation in DNA, which is mostly inherited from parents or it can be caused by certain lifestyle style related risk factors.

Even though we have very high death rate due to breast cancer, effective way to prevent it from occurring has not been yet found. Regular checkups like mammography, breast ultrasound or magnetic resonance imaging every year between age of 45 to 54, every two years after age of 55 and consultation with doctor starting age of 40 can help early detection of breast cancer. Below table shows death estimation due to breast cancer for year. [27]

Table 3 CDC Estimated Female Breast Cancer Cases and Deaths by Age, US, 2017

Age	**InSitu Cases**		**Invasive Cases**		**Deaths**	
	Number	**%**	**Number**	**%**	**Number**	**%**
<40	1,610	3%	11,160	4%	990	2%
40-49	12,440	20%	36,920	15%	3,480	9%
50-59	17,680	28%	58,620	23%	7,590	19%
60-69	17,550	28%	68,070	27%	9,420	23%
70-79	10,370	16%	47,860	19%	8,220	20%
80+	3,760	6%	30,080	12%	10,910	27%
All ages	**63,410**		**252,710**		**40,610**	

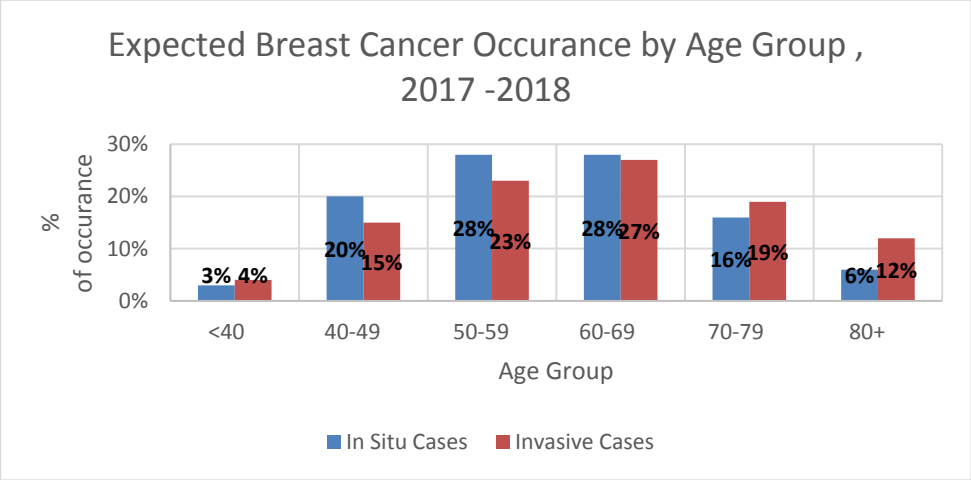

Figure 6 Expected 2017- 2018 Breast Cancer Occurrence by Age Group

From figure 5 we can see that chances of occurrence of breast cancer is more during 50s and trend continues till late 60s.

3.1.3 Diabetes

Diabetes is a condition when body starts to produce too much sugar in the blood. In this condition, body doesn't properly process food to use as energy and most of the food person eats converted into glucose. Pancreas either doesn't make enough insulin to help get glucose into blood or can't use available insulin as it should, this causes sugars to build up in body and results in diabetes.

Diabetes can be of different type, pre-diabetes, type 1 diabetes or type 2 diabetes. Pre-diabetes is the condition when blood sugar is high but not enough to result in type 2 diabetes , type 1 diabetes is condition when pancreases either produces very little insulin or no insulin at all and type 2 diabetes is chronic condition which affects the way body processes blood sugar level . Type 2 diabetes accounts for 90% to 95% of all diabetes cases [28]

Frequent urination, obesity, sudden weight loss, sudden vision changes, numbness in hands or feet, feeling tired most of the times extreme hunger or thrust , dryness in skin and slow healing are most common symptoms in diabetic patient. Blood glucose test is most common way of detecting diabetes.

According to national diabetes statistic report, an estimated 30.3 million people of all ages had diabetes in 2015 out of which 23.8% patients were not aware of having diabetes. Study shows older the age higher the rate of diabetes is. 1.5 million Americans are diagnosed with diabetes every year and it remains the seventh leading cause of death in United State. [29]

Table 4 Estimated Diabetes Adults aged ≥18 years, US, 2015

Characteristic	Diagnosed diabetes No. in millions (95% CI)a	Undiagnosed diabetes No. in millions (95% CI)a	Total diabetes No. in millions (95% CI)a
Total	23.0 (21.1–25.1)	7.2 (6.0–8.6)	30.2 (27.9–32.7)
Age in years			
18–44	3.0 (2.6–3.6)	1.6 (1.1–2.3)	4.6 (3.8–5.5)
45–64	10.7 (9.3–12.2)	3.6 (2.8–4.6)	14.3 (12.7–16.1)
≥65	9.9 (9.0–11.0)	2.1 (1.4–3.0)	12.0 (10.7–13.4)
Sex			
Women	11.7 (10.5–13.1)	3.1 (2.4–4.1)	14.9 (13.5–16.4)
Men	11.3 (10.2–12.4)	4.0 (3.0–5.5)	15.3 (13.8–17.0)
	Percentage	Percentage	Percentage
	(95% CI)b	(95% CI)b	(95% CI)b
Total	9.3 (8.5–10.1)	2.9 (2.4–3.5)	12.2 (11.3–13.2)
Age in years			
18–44 Table	2.6 (2.2–3.1)	1.3 (0.9–2.0)	4.0 (3.3–4.8)
45–64	12.7 (11.1–14.5)	4.3 (3.3–5.5)	17.0 (15.1–19.1)
≥65	20.8 (18.8–23.0)	4.4 (3.1–6.3)	25.2 (22.5–28.1)
Sex			
Women	9.2 (8.2–10.3)	2.5 (1.9–3.2)	11.7 (10.6–12.9)
Men	9.4 (8.5–10.3)		12.7 (11.5–14.1)

Where CI= Confidence interval, a =Numbers for subgroups may not add up to the total because of rounding.

b =Data are crude, not age-adjusted, Data source: 2011–2014 National Health and Nutrition Examination Survey and 2015 U.S. Census Bureau data.

3.2 Logic Extraction and Proposed Model

Health risk adjustment is method of comparing populations and adjusting health plan payments using health status of members and these health status is collected through electronic medical record [41], medical and pharmacy claims. The CDPS, is a risk adjustment system developed explicitly for states to use in adjusting capitated

payments for Medicaid enrollees, uses diagnosis codes to classify enrollees into 19 different condition categories, 18 of which we used to designate someone as having a chronic or disabling condition. The CDPS uses the first three digits of each diagnosis code to classify people into 19 major diagnostic categories. [30]

CDPS- Rx model uses linear regression to calculate risk scores based on inpatient, outpatient diagnosis for chronic conditions of member, member demographic, disabilities and drug prescription. This model excludes codes that are not well defined among clinicians and also excludes many diagnosis codes that are low cost and high frequency of occurrence since these kind of diagnosis do not contribute on patients chronic conditions. CDPS + RX model is one of the predictive model which helps stratifying member's health risk and this algorithm is available in SAS programming language. CDPS+Rx model was developed by University of San Diego. CDPS is a risk adjustment system for Medicaid which maps available less common but costly chronic diagnosis to 58 CDPS categories, these diagnosis are selected based on their occurrence disabled Medicaid beneficiaries and Medicaid Rx model is pharmaceutical based model using NDC codes to assign 45 therapeutic categories. Combined CDPS + Rx model uses 15 MRX categories.

This algorithm has three main steps and below is detail explanation of these three steps. First defining diagnosis hierarchies, this step is built to classify ICD diagnosis codes into CDPS diagnostic categories. Base model stratifies each diagnostic categories into hierarchical levels of severity, as high, medium and low [30]. Level of severity denotes the level of healthcare a patient needs. Each diagnostic code is defined under diagnostic category and level of severity. When patient has more than one

diagnosis for same diagnostic group, diagnosis contributing to highest level is retained and lower levels are assigned weight zero.

Second is grouping NDCs under 15 MRX categories. In this step algorithm uses NDC codes to define them into 15 different categories. It runs logic for categorizing NDCs, 15 NDC MRX categories and using labeled categories to specific conditions.

Third step is to combine weights extracted from member eligibility, diagnosis and drug codes and build a combined diagnosis and pharmacy risk adjustment model by applying normalization factors. For our research purpose we have excluded normalized risk calculation step. We have only calculated individual risk scores for our observation which is also called risk assessment, excluding normalized risk calculation allows us to include all kind of membership including Medicare, as we are not calculating payment method.

In our proposed model we took the diagnostic categories and NDC categories ad their respective weights described by existing model. Figure 4 shows proposed risk and disease management model. Proposed model is implemented in 2 major steps: first step is risk assessment and second step is membership identification for disease management programs once risk score is calculated. In our proposed model based on derived diagnostic and NDC categories we built our diagnosis and drug hierarchy in structured query language and once we validated risk score calculation variable using linear regression algorithm we used same sample of observation with added risk scores and utilization metrics to determine disease management logic. For testing purpose we have only included chronic disease mentioned in section 3.1 in the proposed model.

Proposed model is clearly new model which calculated risk score of individual patients along with providing disease management member eligibility flag. Also proposed model is built in data warehouse language which makes this model productive and efficient in terms of data preparation and model execution time. Detail on proposed model logic validation and model implementation is described in section 3.3 and 3.4 respectively.

3.3 Validation Using Predictive Analysis

We have divided data into training data set and test data with ratio of 70% to 30% respectively.

RSDATA is our data file with 4761 observation for 3 major chronic conditions for state fiscal year 2016. After dividing data into training and testing set we have 3342 observation on training data set and 1419 observations on testing dataset.

Then we applied linear regression model to our training data set, in our case,

$$\text{RISK SCORE} = \text{DEMO} + \text{INTER} + \text{MEDI} + \text{MRX}$$

Based on our training data and extracted logic from base model we applied multiple linear regression equation as above. State Score is our predicted output based on variables,

- DEMO: demographic score
- INTER: Intercept score using healthcare benefit eligibility
- MEDI: Diagnosis weight from medical claims
- MRX: Diagnosis weight from pharmacy claims

Based on output from regression model we predicted our model in test data set and compared final score of training observations to predicted test observations. Once risk score is calculated on test data and accuracy is evaluated, then we took same sample combining with member utilization data and ran fast and frugal decision tree prediction model to identify most contributing variable to identify membership for disease management program. We ran fast and frugal decision tree algorithm on training data which is 70% of overall observation. We combined risk assessment outcome with member's actual disease and case management status and their behavioral and clinical admits and visits information to run this validation. Our model not only ranked which factors are highly contributed on selecting members for disease management programs but also provided performance comparison against other classification model like SVM, CART, LR and RF.

3.4 Model Implementation

We have developed risk assessment model in structured query language based on logic extracted using basic concept of chronic illness and disability payment system and Rx. Implementing algorithm in same database system where data is warehoused makes any analysis and prediction efficient in terms of productivity. It saves time to import or export data or output, if any changes need to be done in the script while selecting data then that can be done without any hassle since everything is stored in same database, along with time this approach saves cost as you only need single platform to implement logic and view result. This is the main reason we have chosen SQL database.

The proposed model is implemented on windows 7 enterprise operating system using features of Microsoft SQL server 2016 and R studio as statistical validation tool. All experiments are implemented on a Dell laptop of Intel Core i5-5200U CPU @ 2.20 GHz with 64 bit operating system and 8.00 GB RAM.

Based on member's plan eligibility and healthcare benefit eligibility we have defined demographic input

We have prepared excel dataset with all the input dataset provided in base model, these dataset categorizes diagnosis hierarchy, drug classification based, weight based on specific category of assistance and imported in our SQL database for further use. Based on diagnosis weight, member demographic, member eligibility and institutional claim's historical data we then selected members claims information.

Based on hierarchy of diagnosis and demographic information we then allocated 0 or 1 variable to each input variable. If members has specific diagnosis, fall under specific age and gender band then it is 1 else 0 , and have allocated true condition to each intercept variable since its based upon member's plan eligibility and its true for all cases.

Similarly, we selected pharmacy claims information from pharmacy data warehouse for the same observation. Now again using pivot function we filtered members who has positive value for at least one of diagnosis or pharmacy code based on our pharmacy grouper. Referring to base model we have created our database for weight related to each diagnosis and drugs.

In the next step we extracted utilization metric like readmission rate, emergency visits, health behaviors such as smoking habit, alcohol consumption, and their current

case management, care-coordination or disease management flags and combined this data with our risk assessment model.

After combining member's calculated risk scores with their medical utilization data, we applied Fast and Frugal Decision Trees (FFDT) testing model to check what components should be used for selecting patients for disease management program. Based on FFDT outcome for our observations we flagged for disease management tier as:

- Tier I: Member has over all high risk score which means member with risk category >1 and risk score >2.15, total inpatient admits and emergency visits > 2 and preventive visit > 1
- Tier II: Medium risk score members with total inpatient and emergency visit <2 and preventive visit >1
- Tier III: All other members

For disease management purpose Tier I members will get priority over tier II since their medical conditions are deteriorating compared to I. So our chronic condition specification is dependent on having diagnosis for either one of these disease or more. Each of these disease has been described in section 3.1 of this chapter and these conditions have been categorized based on diagnosis categorized by ICD9/10 grouper model.

Once patients are risk stratified and flagged for disease management programs these members will be sent to respective departments for further observations in terms of cost and health care. In any healthcare organization finance team can utilize provided information for cost management/forecasting as higher the risk score more will be

spending cost on that member. Operations team can utilize same information for provider management and education purpose and clinical team and physicians can utilize information for care/disease management purpose. Based on member's risk scores and DM flag clinical team can outreach to selected patients, help patient manage necessary medical service and provide necessary medical education to the patients. This will greatly help clinical team in term of time for identifying membership since proposed risk and disease management model automatically prioritize members eligible for disease management and monitoring program and helped them focusing more on quality of care that needs to be provided.

4 Experimental Results and Discussion

4.1 Score Validation Using Linear Regression and Illinois States Risk Score

We didn't expect 100% matching of risk score against state individual score because there with time data cleaning occurs and we have pulled most recent and final data from Molina Inc.'s data warehouse. For validation purpose we ran our observation into proposed SQL based risk and disease management model and selected SAS base model. Though we had final individual scores from state using base model, we wanted to add one more step towards validation of our calculated risk scores so we ran original base Model in SAS. We assume that state risk adjustment model is equivalent to original base SAS model.

First we tested for correlation between state score and medical and drug weights and found that they are linearly related. Weights from medical diagnosis and drug codes in addition to demographic and healthcare eligibility are the key contributor to generate risk scores in the training phase. The trained observations are then evaluated

with test data. Figure 7 shows calculation of scores based on training data and figure 8 shows our scores based on predicted model from training set that we applied on test data.

Applying multiple linear regression on test observation yield 99% of confidence with R-squared value of 98%, which means 98% of selected variables are contributing in calculation of risk score for proposed model and that only 2% of data are not close to fitted regression line.

We received Probability-value (p-value) < 2.2e-16, which tells significantly our input variables are contributing to the proposed model. When is interpret our p-value we get < .00000000000000022, which is much smaller than the conventional value of 0.05 which defines significance of input variable in the model and our output shows that observations in our proposed model are highly significant.

Table 5 Multiple Linear Regression Result on Test Observations

Factors	**Yield**
P-Value	< 2.2e-16
Multiple R-squared	0.9887
Adjusted R-squared	0.9886

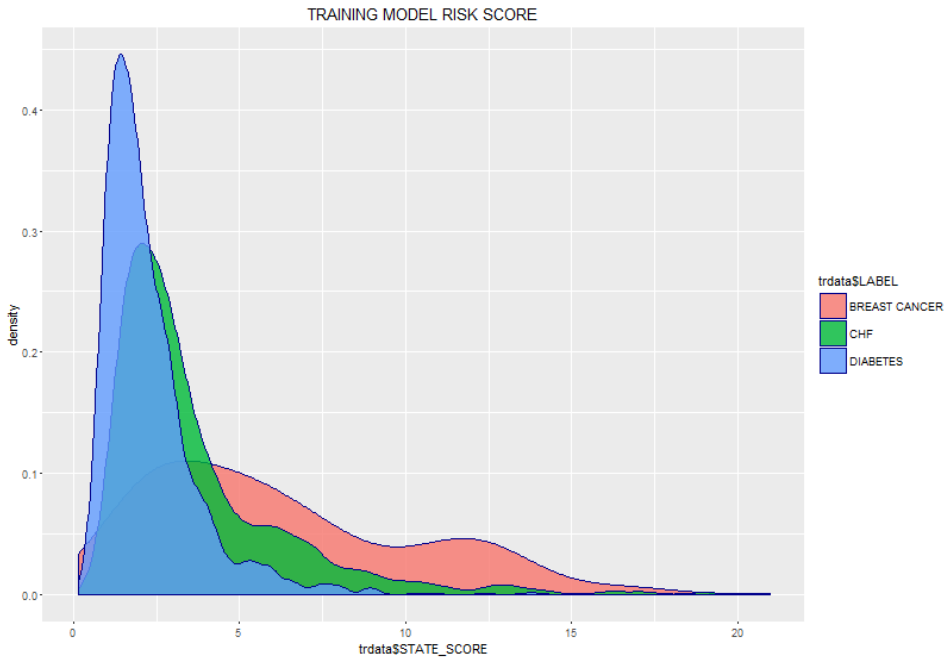

Figure 7 Predicted Outcome on Training Observations

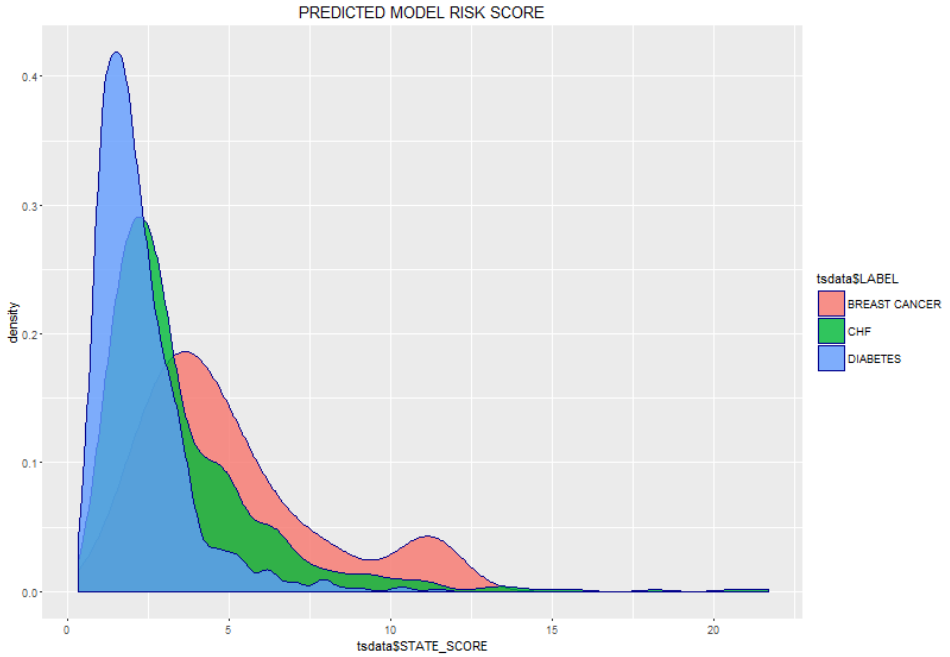

Figure 8 Predicted Outcome on Test Observations

4.2 Fast and Frugal Decision Tree Output for DM Member Selection

After implementing risk assessment model we extracted further medical utilization metric (UM) for the same observations along with their disease management status. Once we combine calculated risk scores and UM metric for our observations we divided data into training and test set as 7:3 ratio respectively and ran fast and frugal decision tree algorithm on R.

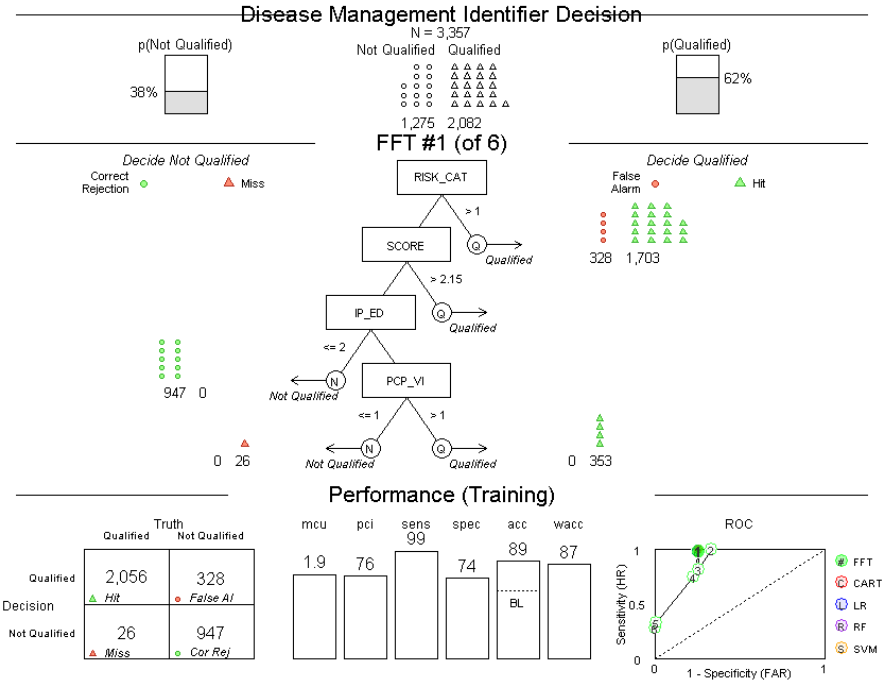

Figure 9 FFDT Outcome on Training Observations

Figure 9 shows that on training dataset we achieved accuracy of 89% with sensitivity of 99% and specificity 74% and discovered that risk score and risk category is highest contributing factor. Our result on training dataset shows that 89% of predicted value are true to actual value which is shown by accuracy, 99% of tested observations that are predicted as positive are actually positive observations which is

shown by sensitivity and 74% of observations that are actually negative are predicted as negative which is shown by specificity.

Then we applied our prediction to test dataset, Figure 10 shows result from test data, our model yield 88% of accuracy which means 89% of predicted value are true to actual value, with 99% sensitivity which means 99% of tested observations that are predicted as positive are actually positive observations and 70% specificity which means 74% of observations that are actually negative are predicted as negative on test dataset.

ROC curve for 1- specificity (proportion of false alarms) vs sensitivity shows performance comparison of FFDT and four other classification trees and our selection of FFDT for testing purpose is correct, Table 6 demonstrates the performance of each tree. We result shows that compared to models SVM, LR, CART, RF , our selected FFTrees algorithm is better in terms of performance. ROC curve is a plot of True Positive (TP) Rates against False Positive (FP) Rates where FP rate is the ratio of false positive results to all negative samples [31]. In figure 11, our training models show that risk scores and risk categories are highest contributing factors for identifying members for disease management program compared to inpatient visit, emergency visits and behavior habits like smoking. Both risk score and risk categories are notation for risk assessment scores calculated from proposed model and both have higher level of sensitivity, accuracy and specificity values.

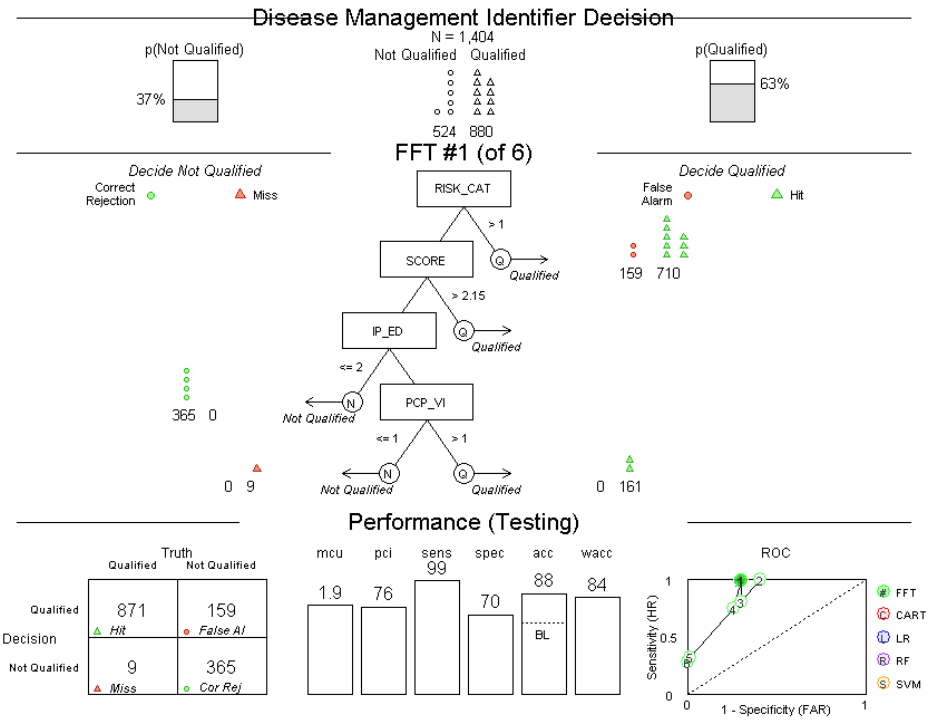

Figure 10 FFDT Outcome on Test Observations

Table 6 FFDT vs Other Classification Tree Output Comparison on Training and Test Observations

OBSERVATION	FFTrees	LR	CART	RF	SVM
Train	89%	87%	87%	88%	87%
Test	87%	84%	85%	85%	85%

Figure 11 Input Variable Contribution for DM Flag

4.3 Result Comparison

After implementing extracted logic in our proposed chronic risk and disease management model, we ran data against original base model and also compared our calculated scores against available scores from state. Table 7 shows execution time for proposed model in structured query language. Proposed model is 2 minutes slower than original model but this difference is considerable since we have added calculation for member identification for disease management programs too. Table 8 and figure 12 shows average risk scores for proposed model, state scores and original model, our scores show that we the variance is minimal which aligns with achieved 99% of accuracy of proposed model.

Table 7 Execution Time Comparison on Each Model

Model	Execution Time (mm:ss)
Proposed	27:34
State	NA
SAS CDPS	25:11

Table 8 Average Calculated Risk Score Using Each Model

	Average Risk Score		
Model	CHF	Breast Cancer	Diabetes
Proposed	3.76	5.97	2.31
State	3.66	5.89	2.27
SAS CDPS	3.67	5.70	2.29

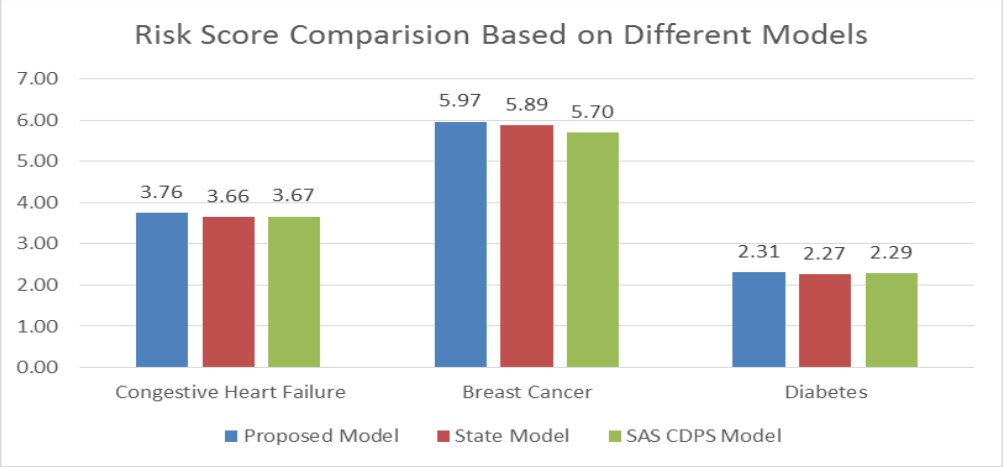

Figure 12 Risk Score Comparison Based on Each Model for 3 Chronic Conditions

Figure 12 is graphical representation of table 8, which shows average risk scores for selected major three chronic conditions for test samples used in proposed model. Risk scores can vary from 0.067 to 39.679 based on member's demographics, eligibility, health conditions and medications they are using. In our observation set our overall calculated lower bound risk score for selected chronic conditions of our total

observations is 0.45 and upper bound is 23.57. Figure 12 and table 8 shows average risk scores for the selected chronic conditions of the observation that were used to implement proposed model compared to state scores and original model, our calculated average risk scores aligns with state model and base model with marginal variance.

Feature comparison of proposed model against original base model and state model is shown in table 9. Our chronic risk and disease management model provides member identification feature for eligible and required patients. We validated this flag against the disease management status for these member we currently have in system which have been identified by professional clinicians and expert physicians. Table 10 and 11 show the confusion matrix for training and test observations that we have achieved from fast and frugal decision tree. Confusion matrix is a matrix representation of the classification result [32].

Table 9 Feature Comparison Base Model vs Proposed Model

	Features	
Model	**Risk Score**	**Disease**
Proposed	Y	Y
State Model	Y	N
CDPS Base	Y	N

Table 10 Confusion Matrix for Training Observations

Result- Train Data	**Based on DM flag Identified by Expert Clinicians & Physicians**	
	TRUE	**FALSE**
Positive	2056 (TP)	328 (FP)
Negative	26 (TN)	947 (FN)

Table 11 Confusion Matrix for Test Observations

Result- Test Data	**Based on DM flag Identified by Expert Clinicians & Physicians**	
	TRUE	**FALSE**
Positive	871 (TP)	159 (FP)
Negative	9 (TN)	365 (FN)

We do have higher number of false positives in both test and train data set but our result from FFDT model shows that risk scores and risk categories are primary contributors for disease management programs. These 487(Train FP 328 + Test FP 159) false positive members could be those un-identified or missed additional high risk members that needs to be considered for disease management programs and this supports in achieving one of the objectives of this research.

5. Conclusion

Population health management is very important and critical sector in healthcare industry, main focus of population health management is to identify sicker member, improve and monitor their clinical condition along with managing financial cost. One of the best ways to improve member health is by identifying chronic and non-chronic members and based on their health status categories members under certain care-coordination or disease management programs and assign case managers. This is done by nurse practitioner or physicians based on one's experience and available medical data and sometime they might miss certain information as medical, pharmacy and member demographic information that comes in various source and in multiple segments. Hence, to support physicians and over all clinical team, need of risk scoring and disease management member identifier system arises.

Risk scoring of each members based on their demographic, medical and pharmacy helps provider groups, clinical teams and health insurance companies to find their risky members in terms of care and cost. In this paper, we proposed calculation of risk scores of patients in the same source where data is located and gets refreshed as soon as new information is received and based on calculated scores we flagged our members for disease management programs. This approach used structure query language to calculate patient's risk score and to flag DM members. We used member's medical record, pharmacy utilization, demographic information and medical claims data is being used to calculate patient's risk score [36] based on existing risk adjustment model [7] and developed an efficient and new risk assessment model with additional feature for identifying members for disease management or for care-coordination. With the help of designed model we not only can track our risky members on real time basis as soon as we receive their medical information but also at the same time we can suggest them to our clinical teams for required care coordination programs or monitoring purpose. Since we are only using risk assessment part of risk adjustment model, we are able to consider all the population. Also looking at risk scores of all the members of certain provider we can verify that if corresponding provider is enrolling both healthy and sicker members or not and how risky our members are in terms of cost and health [37] [6]. For example if we select members of same age for a physician group and their calculated risk score comes out to be same on lower side of risk score or members have very minimal difference with risk score close to lower bound of risk table then we know that particular physician is accepting

members based on their good health conditions, this is called risk selection and proposed model helps us performing such analysis.

To show that proposed method is efficient and accurate we validated our method using multiple linear regression algorithm and validate our calculated scores against available individual risk scores from State of Illinois for same members, as a second validation step we ran original base model. To identify contributing factors in selecting members for disease management we ran fast and frugal decision tree. The results show proposed chronic risk and disease management algorithm not only calculates health risk of patients but also confirms that risk score can be very important contributing factor for identifying members for disease management and monitoring purpose. Using the proposed model we achieved 99% of confidence level for risk assessment and achieved 89% of accuracy with 99% of sensitivity and 74% of specificity on calculated risk scores as highly contributing factors for identifying members for disease management programs.

In summary, using proposed chronic risk and disease management system any healthcare providers or physician groups can track their member's health risk status, based on risk scores identify members for specific care program and track their progress in terms of medical conditions, which ultimately helps member in term of managing their health. Future development of this application can be the following:

- We have used one year worth of data for 3 major chronic conditions. There are total 4761 observations out of which 3342 are for training purpose. In future we can add more chronic conditions for multiple years so that we have wider range for risk scoring purpose.

- In this study we have excluded risk adjustment payment calculation step, in future payment calculation for both Medicaid and Medicare population can be added.
- In future more risk adjustment algorithms should be studied and relevant logic should be extracted from each of them and then implement those logics in current algorithm to make it more efficient and to include all possible factors that can contribute in chronic risk and disease management process.

6. Appendix A: Algorithms

6.1 Data Selection

```
--DROP TABLE #DIABETES
SELECT DISTINCT MEMID, CAST('Diabetes' as varchar(25)) AS CONDITION
INTO #DIABETES
FROM QNXT_PLANDATA_IL.dbo.CLAIM C1 WITH (NOLOCK)
INNER JOIN QNXT_PLANDATA_IL.dbo.CLAIMDETAIL AS CD  WITH (NOLOCK)
      ON C1.CLAIMID = CD.CLAIMID
LEFT JOIN QNXT_PLANDATA_IL.dbo.CLAIMDIAG AS DX1  WITH (NOLOCK)
      ON CD.CLAIMID = DX1.CLAIMID AND DX1.DIAGTYPE IN ('PRIMARY','1',
      'Secondary','Admit' )
WHERE codeid IN
      (SELECT code FROM [Illinois_Report_Details].[dbo].[ALL_DIAG_DEL]
      WITH (NOLOCK)
        WHERE Value_Set_Name IN ('Diabetes'))

AND CAST(startdate AS DATE) BETWEEN '2015-07-01' AND '2016-06-30'
and ((C1.FORMTYPE = '1500'  and CD.location in ('21'))
or (C1.FACILITYCODE+C1.BILLCLASSCODE IN ('11', '12', '18', '41', '42')
      AND C1.FORMTYPE LIKE 'U%'))
GROUP BY MEMID
HAVING count(distinct c1.MEMID+CAST(STARTDATE AS CHAR))>=1

union

SELECT DISTINCT MEMID,CAST('Diabetes' as varchar(25)) AS CONDITION
FROM QNXT_PLANDATA_IL.dbo.CLAIM C1 WITH (NOLOCK)
INNER JOIN QNXT_PLANDATA_IL.dbo.CLAIMDETAIL AS CD  WITH (NOLOCK)
      ON C1.CLAIMID = CD.CLAIMID
LEFT JOIN QNXT_PLANDATA_IL.dbo.CLAIMDIAG AS DX1  WITH (NOLOCK)
      ON CD.CLAIMID = DX1.CLAIMID AND DX1.DIAGTYPE IN ('PRIMARY','1',
      'Secondary','Admit' )
WHERE codeid IN
      (SELECT code FROM [Illinois_Report_Details].[dbo].[ALL_DIAG_DEL]
      WITH (NOLOCK)
        WHERE Value_Set_Name IN ('Diabetes'))

AND CAST(startdate AS DATE) BETWEEN '2015-07-01' AND '2016-06-30'
and ((C1.FORMTYPE = '1500'  and CD.location in ('11','22','23'))
or (C1.FACILITYCODE+C1.BILLCLASSCODE IN ('13', '14', '43', '83', '85',
      '71', '77') AND C1.FORMTYPE LIKE 'U%'))
GROUP BY MEMID
HAVING count(distinct c1.MEMID+CAST(STARTDATE AS CHAR))>=2

--CHF
--DROP TABLE #ChronicHeartFailure
SELECT DISTINCT MEMID, CAST('Chronic Heart Failure' as varchar(25)) AS
      CONDITION
INTO #ChronicHeartFailure
FROM QNXT_PLANDATA_IL.dbo.CLAIM C1 WITH (NOLOCK)
INNER JOIN QNXT_PLANDATA_IL.dbo.CLAIMDETAIL AS CD  WITH (NOLOCK)
      ON C1.CLAIMID = CD.CLAIMID
LEFT JOIN QNXT_PLANDATA_IL.dbo.CLAIMDIAG AS DX1  WITH (NOLOCK)
```

```
      ON CD.CLAIMID = DX1.CLAIMID AND DX1.DIAGTYPE IN ('PRIMARY','1',
      'Secondary','Admit' )
WHERE codeid IN
      (SELECT code FROM [Illinois_Report_Details].[dbo].[ALL_DIAG_DEL]
      WITH (NOLOCK)
        WHERE Value_Set_Name IN ('Chronic Heart Failure'))

AND CAST(startdate AS DATE) BETWEEN '2015-07-01' AND '2016-06-30'
GROUP BY MEMID
HAVING count(distinct c1.MEMID+CAST(STARTDATE AS CHAR))>=1

--BREAST CANCER
--DROP TABLE #BreastCANCER
SELECT DISTINCT MEMID, CAST('Breast Cancer' as varchar(25)) AS
      CONDITION
INTO #BreastCANCER
FROM QNXT_PLANDATA_IL.dbo.CLAIM C1 WITH (NOLOCK)
INNER JOIN QNXT_PLANDATA_IL.dbo.CLAIMDETAIL  AS CD WITH (NOLOCK)
      ON C1.CLAIMID = CD.CLAIMID
LEFT JOIN QNXT_PLANDATA_IL.dbo.CLAIMDIAG  AS DX1 WITH (NOLOCK)
      ON CD.CLAIMID = DX1.CLAIMID AND DX1.DIAGTYPE IN ('PRIMARY','1',
      'Secondary','Admit' )
JOIN [Illinois_Report_Details].[dbo].[ALL_DIAG_DEL]  WITH (NOLOCK)
      ON      codeid = code
WHERE  Value_Set_Name IN ( 'Breast Cancer')
AND CAST(startdate AS DATE) BETWEEN '2015-07-01' AND '2016-06-30'
and ((C1.FORMTYPE = '1500'  and CD.location in ('21'))
or (C1.FACILITYCODE+C1.BILLCLASSCODE IN ('11', '12', '18', '41', '42')
AND C1.FORMTYPE LIKE 'U%'))
GROUP BY MEMID
HAVING count(distinct c1.MEMID+CAST(STARTDATE AS CHAR)) >= 1

UNION

SELECT DISTINCT MEMID, CAST('Breast Cancer' as varchar(25)) AS
      CONDITION
FROM QNXT_PLANDATA_IL.dbo.CLAIM C1 WITH (NOLOCK)
INNER JOIN QNXT_PLANDATA_IL.dbo.CLAIMDETAIL AS CD WITH (NOLOCK)
      ON C1.CLAIMID = CD.CLAIMID
LEFT JOIN QNXT_PLANDATA_IL.dbo.CLAIMDIAG AS DX1  WITH (NOLOCK)
      ON CD.CLAIMID = DX1.CLAIMID AND DX1.DIAGTYPE IN ('PRIMARY','1',
      'Secondary','Admit' )
JOIN [Illinois_Report_Details].[dbo].[ALL_DIAG_DEL]  WITH (NOLOCK)
      ON      codeid = code
WHERE  Value_Set_Name IN ( 'Breast Cancer')
AND CAST(startdate AS DATE) BETWEEN '2015-07-01' AND '2016-06-30'
and ((C1.FORMTYPE = '1500'  and CD.location in ('11','22','23'))
or (C1.FACILITYCODE+C1.BILLCLASSCODE IN ('13', '14', '43', '83', '85',
      '71', '77') AND C1.FORMTYPE LIKE 'U%'))
GROUP BY MEMID
HAVING count(distinct c1.MEMID+CAST(STARTDATE AS CHAR)) >= 2


DROP TABLE #ALL_Chronic_Conditions
Select * INTO #ALL_Chronic_Conditions from #DIABETES
UNION
Select * from #ChronicHeartFailure
```

```
UNION
Select * from #BreastCancer

--INSERT MEMBER INTO MEMBER TABLE FOR RISK SCORING
DROP TABLE MEMBER
SELECT *  INTO MEMBER FROM #ALL_Chronic_Conditions

--Data Selection criteria for member utilization data
--Logic to Identify 6 month Inpatient Data
SELECT memid, COUNT(DISTINCT memid+CAST(CAST(startdate AS DATE) AS
      CHAR)) AS IP_ADMIT_6M, SUM(servunits) AS BED_DAYS_6M
FROM QNXT_PLANDATA_IL.dbo.CLAIM C1 WITH (NOLOCK)
INNER JOIN QNXT_PLANDATA_IL.dbo.CLAIMDETAIL AS CD WITH (NOLOCK)
ON C1.CLAIMID = CD.CLAIMID
WHERE C1.status in ('PAID','QLTYREVIEW', 'PAY', 'PEND', 'PAYHOLD',
      'WAITPAY')
AND CD.STATUS <> 'DENY'
AND C1.FACILITYCODE+C1.BILLCLASSCODE IN ('11', '12', '18', '41', '42')
AND C1.FORMTYPE LIKE 'U%'
AND REVCODE IN ('0100',   '0101',     '0102',     '0103',     '0104',
        '0105',     '0106',
        '0107',     '0108',     '0109',     '0110',     '0111',
        '0112',     '0113',     '0114',
        '0115',     '0116',     '0117',     '0118',     '0119',
        '0120',     '0121',     '0122',
        '0123',     '0124',     '0125',     '0126',     '0127',
        '0128',     '0129',     '0130',
        '0131',     '0132',     '0133',     '0134',     '0135',
        '0136',     '0137',     '0138',
        '0139',     '0140',     '0141',     '0142',     '0143',
        '0144',     '0145',     '0146',
        '0147',     '0148',     '0149',     '0150',     '0151',
        '0152',     '0153',     '0154',
        '0155',     '0156',     '0157',     '0158',     '0159',
        '0160',     '0161',     '0162',
        '0163',     '0164',     '0165',     '0166',     '0167',
        '0168',     '0169',     '0170',
        '0171',     '0172',     '0173',     '0174',     '0175',
        '0176',     '0177',     '0178',
        '0179',     '0180',     '0181',     '0182',     '0183',
        '0184',     '0185',     '0186',
        '0187',     '0188',     '0189',     '0190',     '0191',
        '0192',     '0193',     '0194',
        '0195',     '0196',     '0197',     '0198',     '0199',
        '0200',     '0201',     '0202',
        '0203',     '0204',     '0205',     '0206',     '0207',
        '0208',     '0209',     '0210',
        '0211',     '0212',     '0213',     '0214',     '0215',
        '0216',     '0217',     '0218',     '0219')
AND CAST(startdate AS DATE) >= dateadd(day,-180,getdate())
GROUP BY      memid
--Logic to Identify 6 month Ememrgency Visit Data
SELECT memid
, COUNT(DISTINCT case when REVCODE IN ('0450', '0451', '0452',
      '0456','0459','0981')
      then memid+CAST(CAST(startdate AS DATE) AS CHAR) else null end)
      AS ED_VISIT_6M
```

```
FROM QNXT_PLANDATA_IL.dbo.CLAIM C1 WITH (NOLOCK)
INNER JOIN QNXT_PLANDATA_IL.dbo.CLAIMDETAIL AS CD WITH (NOLOCK)
      ON C1.CLAIMID = CD.CLAIMID
WHERE C1.status in ('PAID','QLTYREVIEW', 'PAY', 'PEND', 'PAYHOLD',
      'WAITPAY')
AND CD.STATUS <> 'DENY'
AND (
        ( REVCODE IN ('0450', '0451', '0452', '0456','0459','0981')
        OR (C1.FORMTYPE = '1500'  AND CD.location ='23')
        )
      )
AND CAST(startdate AS DATE) >= dateadd(day,-180,getdate())
GROUP BY      memid) ED
ON A.MEMID = ED.MEMID

--Logic to Identify 6 month Preventive Visit Data
SELECT memid, COUNT(DISTINCT memid+CAST(CAST(startdate AS DATE) AS
      CHAR)) AS PREVENTIVE_VISIT_6M
FROM QNXT_PLANDATA_IL.dbo.CLAIM C1
INNER JOIN QNXT_PLANDATA_IL.dbo.CLAIMDETAIL AS CD ON C1.CLAIMID =
      CD.CLAIMID
WHERE C1.status in ('PAID','QLTYREVIEW', 'PAY', 'PEND', 'PAYHOLD',
      'WAITPAY')
AND CD.STATUS <> 'DENY'
AND ( servcode IN (SELECT DISTINCT CODE FROM
      [Illinois_Report_Details].[dbo].[IL_DIM_HEDIS_CODES]WHERE
      Value_Set_Name IN ('Ambulatory Visits', 'Other Ambulatory
      Visits'))
OR billservcode IN (SELECT DISTINCT CODE FROM
      [Illinois_Report_Details].[dbo].[IL_DIM_HEDIS_CODES]WHERE
      Value_Set_Name IN ('Ambulatory Visits', 'Other Ambulatory
      Visits')))
--and CD.location = '11'
AND CAST(startdate AS DATE) >= dateadd(day,-180,getdate())
GROUP BY memid
```

6.2 Logic Extraction

```
* Age and sex variables;
a_under1=(age <= 1);
a_1_4=(1 < age < 5);
a_5_14m=((5 <= age < 15) and male=1);
a_5_14f=((5 <= age < 15) and male=0);
a_15_24m=((15 <= age < 25) and male=1);
a_15_24f=((15 <= age < 25) and male=0);
a_25_44m=((25 <= age < 45) and male=1);
a_25_44f=((25 <= age < 45) and male=0);
a_45_64m=((45 <= age < 65) and male=1);
a_45_64f=((45 <= age < 65) and male=0);
a_65=(65 <= age);
label a_under1="age<=1"
a_1_4="1<age<5"
a_5_14m="5<age<15 male"
a_5_14f="5<age<15 female"
a_15_24m="15<=age<25 male"
```

```
a_15_24f="15<=age<25 female"
a_25_44m="25<=age<45 male"
a_25_44f="25<=age<45 female"
a_45_64m="45<=age<65 male"
a_45_64f="45<=age<65 female"
a_65="65<=age";
run;

data step2(compress=yes);
merge diagind(in=inind)
who(in=inreg);
by recipno;
if inreg;
array dind{*} 3 %names;
retain %names;
length i 3;
drop i;
do i=1 to dim(dind);
if dind{i}=. then dind{i}=0;
end;

* Create hierarchy ;
array vars1{*}  AIDSH INFH HIVM INFM INFL;
array vars2{*}  CANVH CANH CANM CANL;
array vars3{*}  CARVH CARM CARL CAREL;
array vars4{*}  CERL;
array vars5{*}  CNSH CNSM CNSL;
array vars6{*}  DIA1H DIA1M DIA2M DIA2L ;
array vars7{*}  DDM DDL;
array vars8{*}  EYEL EYEVL;
array vars9{*} GENEL;
array vars10{*} GIH GIM GIL;
array vars11{*} HEMEH HEMVH HEMM HEML;
array vars12{*} METH METM METVL;
array vars13{*} PRGCMP PRGINC;
array vars14{*} PSYH PSYM PSYML PSYL;
array vars15{*} SUBL SUBVL;
array vars16{*} PULVH PULH PULM PULL;
array vars17{*} RENEH RENVH RENM RENL;
array vars18{*} SKCM SKCL SKCVL;
array vars19{*} SKNH SKNL SKNVL;

%macro varnum(num);
do i=1 to dim(vars&num)-1;
if vars&num(i)=1 then do j=i+1 to dim(vars&num);
vars&num(j)=0;  end;
end; drop j;
%mend varnum;

%do num=1 %to 19;
%varnum(&num);
%end;

%if &aid=AA or &aid=AC %then %do;
PULH=max(PULVH,PULH); PULVH=0;
%end;
```

```
%if &aid=AA %then %do;
ddl=max(DDM,DDL); ddm=0;
%end;

%if &aid=DC or &aid=AC %then %do;
DIA2L=max(DIA1H,DIA1M,DIA2M,DIA2L);
EYEVL=max(EYEL,EYEVL);
array zerovars{*} DIA1H DIA1M DIA2M EYEL;
do i=1 to dim(zerovars); zerovars{i}=0; end;
%end;

NOCDPS=sum(of %reglst)=0;

* Create interaction variables for the Disabled;
%if &aid=DA %then %do;
array intvars(*) CCARVH CCARM CCNSH CPULVH CPULH CGIH CMETH CHIVM
      CINFM CHEMEH;
do i=1 to dim(intvars); intvars{i}=0; end;
%end;
%else %if &aid=DC %then %do;
array intvars(*) CCARVH CCARM CCNSH CPULVH CPULH CGIH CMETH CHIVM
      CINFM CHEMEH;
array orgvars(*) CARVH CARM CNSH PULVH PULH GIH METH HIVM INFM HEMEH;
do i=1 to dim(intvars); intvars{i}=orgvars{i}; end;
%end;
%if &aid=DA or &aid=DC %then %do;
label CCARVH = 'Childrens CARVH'
CCARM  = 'Childrens CARM'
CCNSH  = 'Childrens CNSH'
CPULVH = 'Childrens PULVH'
CPULH  = 'Childrens PULH'
CGIH   = 'Childrens GIH'
CMETH  = 'Childrens METH'
CHIVM  = 'Childrens HIVM'
CINFM  = 'Childrens INFM'
CHEMEH = 'Childrens HEMEH';
%end;

* CDPS category labels;
%include labfile;
run;
%mend cdps;

%macro mrx;
* Define diagnosis indicator variables;
array dind{*} 3 %names None Other;
retain %names None Other;

if first.recipno  then do;
    do i=1 to dim(dind);
        dind{i}=0;
    end;
end;

* Indicate the existence of a diagnosis in the group variable;

    stage11=put(ndc,$sgrpfmt.);
```

```
    stage21=put(stage11,$snfmt.);
    dind{stage21}=1;

if last.recipno then do;
output;
end;
keep recipno %names;
run;

data who;
set inelig;
run;

data step2(compress=yes);
merge diagind(in=inind)
who(in=inreg);
by recipno;
if inreg;
array dind{*} 3 %names;
retain %names;
length i 3;
drop i;
do i=1 to dim(dind);
if dind{i}=. then dind{i}=0;
end;

* Create hierarchies;
if mrx9=1 then do; mrx8=0; mrx7=0; end;
if mrx8=1 then mrx7=0;

* MRX category labels;
%include labfile;
run;
%mend mrx;

*Combining diagnosis and drug code
if sum(of CARVH CARM)>0 then do; MRX1=0; MRX2=0; end; if MRX1=1 then
      do; CARL=0; CAREL=0; MRX2=0; end;
if sum(of CARL CAREL)>0 then MRX2=0;
if sum(of PSYH PSYM PSYML PSYL)>0 then MRX3=0;
if sum(of DIA1H DIA1M DIA2M DIA2L)>0 then MRX4=0;
if sum(of RENEH RENVH)>0 then MRX5=0; if MRX5=1 then do; RENM=0;
      RENL=0; end;
if HEMEH=1 then MRX6=0; if MRX6=1 then do; HEMVH=0; HEMM=0; HEML=0;
      end;
if sum(of AIDSH INFH)>0 then do; MRX9=0; MRX8=0; MRX7=0; end; if
      MRX9=1 then HIVM=0;
if HIVM=1 then do; MRX8=0; MRX7=0; end; if sum(of MRX7 MRX8 MRX9)>0
      then do; INFM=0; INFL=0; end;
if sum(of SKCM SKCL SKCVL)>0 then MRX10=0;
if sum(of CANVH CANH CANM)>0 then MRX11=0; if MRX11=1 then CANL=0;
if sum(of CNSH CNSM)>0 then do; MRX12=0; MRX13=0; MRX14=0; end; if
      MRX12=1 then do; CNSL=0; MRX13=0; MRX14=0; end;
if CNSL=1 then do; MRX13=0; MRX14=0; end; if MRX14=1 then MRX13=0;
if sum(of PULVH PULH PULM PULL)>0 then MRX15=0;
```

```
if aidcat='DC' then do; if MRX1=1 then CCARM=1; if MRX6=1 then
      CHEMEH=1; if MRX9=1 then do; CHIVM=0; CINFM=0; end; if sum(of
      MRX7 MRX8)=1 then do; CHIVM=1; CINFM=0; end; end;
```

6.3 Validation Using Predictive Analysis

```
#SCRIPT FOR MULTIPLE LINEAR REGRESSION AND SCORE PREDCITION IN TEST
      DATA USING TRAINING DATA AND STATE SCORE

#Multiple Linear Regression
#result<-lm(STATE_SCORE~DEMO+INTER+MEDI+MRX,RSDATA)
#result
#summary(result)

#split dataset into "training" (70%) and "test" (30%)
ind<-sample(2,nrow(RSDATA),replace=TRUE, prob=c(0.7,0.3))
trdata<-RSDATA[ind==1,]
tsdata<-RSDATA[ind==2,]

head(trdata)
head(tsdata)

#Multiple Linear Regression
result<-lm(STATE_SCORE~DEMO+INTER+MEDI+MRX,trdata)
result
summary(result)
coef(result)

#prediction
pred<-predict(result,tsdata)
head(pred)
head(tsdata)

#Visualization Script
# packages
library(dplyr)
library(ggplot2)
library(choroplethr)
library(choroplethrMaps)
library(openintro)
library(diseasemapping)
library(ColorPalette)

head(RSDATA)

#density plot
ggplot(data=RSDATA, aes(RSDATA$STATE_SCORE,fill=RSDATA$LABEL)) +
geom_density(alpha=0.8, color='dark blue')+
ggtitle('STATE RISK SCORE')
#facet_wrap(~RSDATA$LABEL)

ggplot(data=RSDATA, aes(RSDATA$SCORE,fill=RSDATA$LABEL)) +
  geom_density(alpha=0.8, color='dark blue')+
  ggtitle('PREDICTTED RISK SCORE')
```

```
--FFDT Prediction

library(FFTrees)

str(dm_data)

#SPLIT DATASET
set.seed(1234)
ind<- sample(2, nrow(dm_data),replace=T, prob=c(0.7, 0.3))
train <- dm_data[ind==1,]
test <- dm_data[ind==2,]


#STOREING TREE MODEL IN TREE

tree <- FFTrees(formula = CL_DM_FLAG ~ .,
                data = train,
                data.test = test,
                main = "Disease Management Identifier Decision",
                decision.labels = c("Not Qualified", "Qualified")

                )
# to remove comparision with other model do.comp=FALSE

plot(tree, data="train")
plot(tree, data="test")

inwords(tree)
summary(tree)
names(tree)
#for area under curve for both training and test dataset
tree$auc
# to see comparative model  for all samples
tree$decision
cbind(train, tree$decision$train)

# PLOT TREE
plot(tree)
plot(tree, what ='cues')
plot(tree, stats= F)

#tree for test data
plot(tree, data="test")
predict (tree, test)
```

6.4 Model Implementation

```
SELECT DISTINCT
A.MEMID
,A.MEDICAID_ID
,CASE WHEN EK_LOB IN ('ICP','MMP') AND A.Age<18 THEN 'DC'
   WHEN EK_LOB IN ('ICP','MMP') AND A.Age>=18 THEN 'DA'
```

```
              WHEN EK_LOB IN ('FHP','ACA') AND A.Age<18 THEN 'AC'
              WHEN EK_LOB IN ('FHP','ACA') AND A.Age>=18 THEN 'AA'
              END AS AID
,CASE WHEN A.Age <= 1 THEN 'a_under1'
        WHEN A.Age<5 THEN 'a_1_4'
        WHEN A.Age<15 and [sex] = 'M' THEN 'a_5_14m'
        WHEN A.Age<15 and [sex] = 'F' THEN 'a_5_14f'
        WHEN A.Age<25 and [sex] = 'M' THEN 'a_15_24m'
        WHEN A.Age<25 and [sex] = 'F' THEN 'a_15_24f'
        WHEN A.Age<45 and [sex] = 'M' THEN 'a_25_44m'
        WHEN A.Age<45 and [sex] = 'F' THEN 'a_25_44f'
        WHEN A.Age<65 and [sex] = 'M' THEN 'a_45_64m'
        WHEN A.Age<65 and [sex] = 'F' THEN 'a_45_64f'
        WHEN 65 <= A.Age THEN 'a_65'
        END AS Age_GENDER
INTO #MBRSHIP
FROM ILLINOIS_REPORT_DETAILS.DBO.IL_FT_MEMBERMONTHS A
JOIN SFY2016_RS B ON A.MEMID=B.MEMID --2016 State Risk Score
JOIN MEMBER C ON C.MEMID=B.MEMID --Selected 3 chronic Condition
     Membership

SELECT DISTINCT
C1.MEMID
,AID
,B.CATEGORY
INTO #CDPSCLAIMS
FROM QNXT_PLANDATA_IL.dbo.CLAIM C1 WITH (NOLOCK)
  INNER JOIN QNXT_PLANDATA_IL.dbo.CLAIMDETAIL AS CD WITH (NOLOCK)
       ON C1.CLAIMID = CD.CLAIMID
  LEFT JOIN QNXT_PLANDATA_IL.dbo.CLAIMDIAG AS DX1 WITH (NOLOCK)
       ON CD.CLAIMID = DX1.CLAIMID
  LEFT OUTER JOIN QNXT_PLANDATA_IL.dbo.DIAGCODE AS ICD1 WITH (NOLOCK)
       ON DX1.CODEID = ICD1.CODEID
         AND cd.dosfrom between icd1.effdate and icd1.termdate
  JOIN [Illinois_Report_Details].[dbo].[IL_DIM_REF_GROUPER_MASTER] b
      WITH (NOLOCK)
             on REPLACE(dx1.codeid, '.','') = b.code
           and b.model = 'CDPS6.2.2'
  JOIN #MBRSHIP ON #MBRSHIP.MEMID = C1.MEMID

WHERE CAST(startdate AS DATE) BETWEEN '2015-07-01' AND '2016-06-30'
AND DIAGTYPE NOT IN ('PRV','Admit','Trauma')
AND sequence <= 6
AND (formtype like '%UB%'
     or CD.location IN ('03','04','11', '12', '20','22', '23', '21',
     '50', '51','52','53', '54','55', '56', '57', '65', '71', '72'))
ORDER BY 1

SELECT
a.[RecipientID] as medicaid_id
,MEMID
,AID
, DiagCd
INTO #HXCLAIMS1
  FROM [DC01DSSDBPC10\SQL10].[ERR_IL].[hds].[MainClaims] A WITH
      (NOLOCK)
```

```
      JOIN [DC01DSSDBPC10\SQL10].[ERR_IL].[hds].[Institutional] B WITH
      (NOLOCK)
            ON A.DCN = B.DCN
            AND A.[ServiceLineNbr] = B.[ServiceLineNbr]
            AND A.RecipientID = B.RecipientID
            AND A.AdjudicatedDt = B.AdjudicatedDt
      JOIN [DC01DSSDBPC10\SQL10].[ERR_IL].[hds].[Diagnosis] C WITH
      (NOLOCK)
            ON A.DCN = C.DCN
            AND A.[ServiceLineNbr] = C.[ServiceLineNbr]
            AND A.RecipientID = C.RecipientID
            AND A.AdjudicatedDt = C.AdjudicatedDt
       JOIN #MBRSHIP ON #MBRSHIP.MEDICAID_ID = a.[RecipientID]
  WHERE A.RejectionStatusCd = 'N'
  AND [ServiceFromDt] BETWEEN '2015-07-01' AND '2016-06-30'
  GROUP BY a.[RecipientID]
      ,MEMID
      ,AID
      , DiagCd

SELECT
       a.[RecipientID] as medicaid_id
       ,MEMID
      ,AID
      , DiagCd
INTO #HXCLAIMS2
  FROM [DC01DSSDBPC10\SQL10].[ERR_IL].[hds].[MainClaims] A WITH
      (NOLOCK)
      JOIN [DC01DSSDBPC10\SQL10].[ERR_IL].[hds].[Diagnosis] C WITH
      (NOLOCK)
            ON A.DCN = C.DCN
            AND A.[ServiceLineNbr] = C.[ServiceLineNbr]
            AND A.RecipientID = C.RecipientID
            AND A.AdjudicatedDt = C.AdjudicatedDt
      JOIN [DC01DSSDBPC10\SQL10].[ERR_IL].[hds].[NIPS] D WITH (NOLOCK)
            ON A.DCN = D.DCN
            AND A.[ServiceLineNbr] = D.[ServiceLineNbr]
            AND A.RecipientID = D.RecipientID
            AND A.AdjudicatedDt = D.AdjudicatedDt
     JOIN #MBRSHIP ON #MBRSHIP.MEDICAID_ID = a.[RecipientID]
  WHERE A.RejectionStatusCd = 'N'
  AND [ServiceFromDt] BETWEEN '2015-07-01' AND '2016-06-30'
  AND [PlaceOfServiceCd] IN ('03','04','11', '12', '20','22', '23',
      '21', '50', '51','52','53', '54','55', '56', '57', '65', '71',
      '72', 'A', 'E', 'B', 'C', 'G')
  GROUP BY a.[RecipientID]
       ,MEMID
       ,AID
      , DiagCd

SELECT DISTINCT
       a.[RecipientID] as medicaid_id
       ,MEMID
       ,AID
      , DiagCd
       INTO #HXCLAIMS3
  FROM [DC01DSSDBPC10\SQL10].[ERR_IL].[hds].[MMP_MainClaims] A
```

```
      JOIN [DC01DSSDBPC10\SQL10].[ERR_IL].[hds].[MMP_Institutional] B
            ON A.DCN = B.DCN
            AND A.[ServiceLineNbr] = B.[ServiceLineNbr]
            AND A.RecipientID = B.RecipientID
            AND A.AdjudicatedDt = B.AdjudicatedDt
      JOIN [DC01DSSDBPC10\SQL10].[ERR_IL].[hds].[MMP_Diagnosis] C
            ON A.DCN = C.DCN
            AND A.[ServiceLineNbr] = C.[ServiceLineNbr]
            AND A.RecipientID = C.RecipientID
            AND A.AdjudicatedDt = C.AdjudicatedDt
       JOIN #MBRSHIP ON #MBRSHIP.MEDICAID_ID = a.[RecipientID]
  WHERE A.RejectionStatusCd = 'N'
  AND [ServiceFromDt] BETWEEN '2015-07-01' AND '2016-06-30'
  GROUP BY a.[RecipientID]
      ,MEMID
      ,AID
      , DiagCd

 SELECT DISTINCT
        a.[RecipientID] as medicaid_id
       ,MEMID
       ,AID
      , DiagCd
INTO #HXCLAIMS4
  FROM [DC01DSSDBPC10\SQL10].[ERR_IL].[hds].[MMP_MainClaims] A
      JOIN [DC01DSSDBPC10\SQL10].[ERR_IL].[hds].[MMP_Diagnosis] C
            ON A.DCN = C.DCN
            AND A.[ServiceLineNbr] = C.[ServiceLineNbr]
            AND A.RecipientID = C.RecipientID
            AND A.AdjudicatedDt = C.AdjudicatedDt
      JOIN [DC01DSSDBPC10\SQL10].[ERR_IL].[hds].[MMP_NIPS] D
            ON A.DCN = D.DCN
            AND A.[ServiceLineNbr] = D.[ServiceLineNbr]
            AND A.RecipientID = D.RecipientID
            AND A.AdjudicatedDt = D.AdjudicatedDt
      JOIN #MBRSHIP ON #MBRSHIP.MEDICAID_ID = a.[RecipientID]
  WHERE A.RejectionStatusCd = 'N'
  AND [ServiceFromDt] BETWEEN '2015-07-01' AND '2016-06-30'
  AND [PlaceOfServiceCd] IN ('03','04','11', '12', '20','22', '23',
      '21', '50', '51','52','53', '54','55', '56', '57', '65', '71',
      '72', 'A', 'E', 'B', 'C', 'G')
  GROUP BY a.[RecipientID]
      ,MEMID
      ,AID
      , DiagCd


SELECT * INTO #ALL_HXCLAIMS FROM #HXCLAIMS1
UNION
SELECT * FROM #HXCLAIMS2
UNION
SELECT * FROM #HXCLAIMS3
UNION
SELECT * FROM #HXCLAIMS4

SELECT * FROM #ALL_HXCLAIMS
WHERE MEMID IN
```

```
(
      SELECT distinct A.MEMID FROM
      [Illinois_Report_Details].[dbo].MEMBER A
      JOIN
        SFY2016_RS B ON A.MEMID=B.MEMID
)

SELECT DISTINCT MEMID, AID, 'Intercept' as CATEGORY
INTO #CDPSCLAIMS2
FROM #MBRSHIP A
UNION
SELECT DISTINCT MEMID, AID, AGE_GENDER as CATEGORY
FROM #MBRSHIP A
UNION
SELECT DISTINCT MEMID, AID, CATEGORY
FROM #CDPSCLAIMS A
UNION
SELECT DISTINCT MEMID, AID, CATEGORY
FROM #HXCLAIMS1 A
LEFT JOIN [Illinois_Report_Details].[dbo].[IL_DIM_REF_GROUPER_MASTER]
      b
        on DiagCd = b.code
       and b.model = 'CDPS6.2.2'
UNION
SELECT DISTINCT MEMID, AID, CATEGORY
FROM #HXCLAIMS2 A
left JOIN [Illinois_Report_Details].[dbo].[IL_DIM_REF_GROUPER_MASTER]
      b
        on DiagCd = b.code
        and b.model = 'CDPS6.2.2'
UNION
SELECT DISTINCT MEMID, AID, CATEGORY
FROM #HXCLAIMS3 A
LEFT JOIN [Illinois_Report_Details].[dbo].[IL_DIM_REF_GROUPER_MASTER]
      b
        on DiagCd = b.code
        and b.model = 'CDPS6.2.2'
UNION
SELECT DISTINCT MEMID, AID, CATEGORY
FROM #HXCLAIMS4 A
LEFT JOIN [Illinois_Report_Details].[dbo].[IL_DIM_REF_GROUPER_MASTER]
      b
        on DiagCd = b.code
        and b.model = 'CDPS6.2.2'

SELECT MEMID, AID,
      ISNULL(Intercept,0)Intercept,a_under1,a_1_4,a_5_14m,a_5_14f,a_15
      _24m,a_15_24f,a_25_44m,a_25_44f,a_45_64m,a_45_64f,a_65,isnull([C
      ARVH],0)[CARVH], isnull([CARM],0)[CARM], isnull([CARL],0)[CARL],
      isnull([CAREL],0)[CAREL],
isnull([PSYH],0)[PSYH], isnull([PSYM],0)[PSYM],
      isnull([PSYML],0)[PSYML], isnull([PSYL],0)[PSYL],
      isnull([SKCM],0)[SKCM],
isnull([SKCL],0)[SKCL], isnull([SKCVL],0)[SKCVL],
      isnull([CNSH],0)[CNSH], isnull([CNSM],0)[CNSM],
      isnull([CNSL],0)[CNSL],
```

```
isnull([PULVH],0)[PULVH], isnull([PULH],0)[PULH],
      isnull([PULM],0)[PULM], isnull([PULL],0)[PULL],
      isnull([GIH],0)[GIH],
isnull([GIM],0)[GIM], isnull([GIL],0)[GIL], isnull([DIA1H],0)[DIA1H],
      isnull([DIA1M],0)[DIA1M], isnull([DIA2M],0)[DIA2M],
isnull([DIA2L],0)[DIA2L], isnull([SKNH],0)[SKNH],
      isnull([SKNL],0)[SKNL], isnull([SKNVL],0)[SKNVL],
      isnull([RENEH],0)[RENEH],
 isnull([RENVH],0)[RENVH], isnull([RENM],0)[RENM],
      isnull([RENL],0)[RENL], isnull([SUBL],0)[SUBL],
      isnull([SUBVL],0)[SUBVL],
 isnull([CANVH],0)[CANVH], isnull([CANH],0)[CANH],
      isnull([CANM],0)[CANM], isnull([CANL],0)[CANL],
      isnull([HLTRNS],0)[HLTRNS],isnull([DDM],0)[DDM],
 isnull([DDL],0)[DDL], isnull([GENEL],0)[GENEL],
      isnull([METH],0)[METH], isnull([METM],0)[METM],
      isnull([METVL],0)[METVL],
  isnull([PRGCMP],0)[PRGCMP], isnull([PRGINC],0)[PRGINC],
      isnull([EYEL],0)[EYEL], isnull([EYEVL],0)[EYEVL],
      isnull([CERL],0)[CERL],
   isnull([AIDSH],0)[AIDSH], isnull([INFH],0)[INFH],
      isnull([HIVM],0)[HIVM], isnull([INFM],0)[INFM],
      isnull([INFL],0)[INFL],
   isnull([HEMEH],0)[HEMEH], isnull([HEMVH],0)[HEMVH],
      isnull([HEMM],0)[HEMM], isnull([HEML],0)[HEML]
INTO #DIAGIND
FROM
(SELECT distinct memid,AID, CATEGORY FROM #CDPSCLAIMS2) AS s
PIVOT(count(CATEGORY) FOR CATEGORY IN
      (Intercept,a_under1,a_1_4,a_5_14m,a_5_14f,a_15_24m,a_15_24f,a_25
      _44m,a_25_44f,a_45_64m,a_45_64f,a_65,[CARVH],[CARM],[CARL],[CARE
      L],[PSYH],[PSYM],[PSYML],[PSYL],[SKCM],[SKCL],[SKCVL],[CNSH],[CN
      SM],
[CNSL],[PULVH],[PULH],[PULM],[PULL],[GIH],[GIM],[GIL],[DIA1H],[DIA1M],
      [DIA2M],[DIA2L],[SKNH],[SKNL],[SKNVL],[RENEH],[RENVH],[RENM],
[RENL],[SUBL],[SUBVL],[CANVH],[CANH],[CANM],[CANL],[HLTRNS],[DDM],[DDL
      ],[GENEL],[METH],[METM],[METVL],[PRGCMP],[PRGINC],[EYEL],[EYEVL]
      ,[CERL],
[AIDSH],[INFH],[HIVM],[INFM],[INFL],[HEMEH],[HEMVH],[HEMM],[HEML])  )
      pvt

update #Diagind set CARVH=1, PULVH=1 where HLTRNS=1
update #Diagind set INFH=0,HIVM=0,INFM=0,INFL=0 where AIDSH=1
update #Diagind set HIVM=0,INFM=0,INFL=0 where INFH=1
update #Diagind set INFM=0,INFL=0 where HIVM=1
update #Diagind set INFL=0 where INFM=1
update #Diagind set CANH=0,CANM=0,CANL=0 where CANVH=1
update #Diagind set CANM=0,CANL=0 where CANH=1
update #Diagind set CANL=0 where CANM=1
update #Diagind set CARM=0,CARL=0,CAREL=0 where CARVH=1
update #Diagind set CARL=0,CAREL=0 where CARM=1
update #Diagind set CAREL=0 where CARL=1
update #Diagind set CNSM=0,CNSL=0 where CNSH=1
update #Diagind set CNSL=0 where CNSM=1
update #Diagind set DIA1M=0,DIA2M=0,DIA2L=0 where DIA1H=1
update #Diagind set DIA2M=0,DIA2L=0 where DIA1M=1
update #Diagind set DIA2L=0 where DIA2M=1
```

```
update #Diagind set DDL=0 where DDM=1
update #Diagind set EYEVL=0 where EYEL=1
update #Diagind set GIM=0,GIL=0 where GIH=1
update #Diagind set GIL=0 where GIM=1
update #Diagind set HEMVH=0,HEMM=0,HEML=0 where HEMEH=1
update #Diagind set HEMM=0,HEML=0 where HEMVH=1
update #Diagind set HEML=0 where HEMM=1
update #Diagind set METM=0,METVL=0 where METH=1
update #Diagind set METVL=0 where METM=1
update #Diagind set PRGINC=0 where PRGCMP=1
update #Diagind set PSYM=0,PSYML=0,PSYL=0 where PSYH=1
update #Diagind set PSYML=0,PSYL=0 where PSYM=1
update #Diagind set PSYL=0 where PSYML=1
update #Diagind set SUBVL=0 where SUBL=1
update #Diagind set PULH=0,PULM=0,PULL=0 where PULVH=1
update #Diagind set PULM=0,PULL=0 where PULH=1
update #Diagind set PULL=0 where PULM=1
update #Diagind set RENVH=0,RENM=0,RENL=0 where RENEH=1
update #Diagind set RENM=0,RENL=0 where RENVH=1
update #Diagind set RENL=0 where RENM=1
update #Diagind set SKCL=0,SKCVL=0 where SKCM=1
update #Diagind set SKCVL=0 where SKCL=1
update #Diagind set SKNL=0,SKNVL=0 where SKNH=1
update #Diagind set SKNVL=0 where SKNL=1
update #Diagind set PULH=1, PULVH=0  where  (AID='AA' or AID='AC' or
      AID='AG') and PULVH=1 ;
update #Diagind set DDL=1, DDM=0  where  (AID='AA' or  AID='AG') and
      DDM=1 ;
update #Diagind set DIA1H=0, DIA1M=0, DIA2M=0, EYEL=0  where
      (AID='DC' or  AID='AC');

SELECT DISTINCT * ,
CCARVH= case when AID='DC' then CARVH  else 0 end,
CCARM= case when AID='DC' then CARM else 0 end,
CCNSH= case when AID='DC' then CNSH else 0 end,
CPULVH= case when AID='DC' then PULVH  else 0 end,
CPULH= case when AID='DC' then PULH     else 0 end,
CGIH= case when AID='DC' then GIH  else 0 end,
CMETH= case when AID='DC' then METH else 0 end,
CHIVM= case when AID='DC' then HIVM  else 0 end,
CINFM= case when AID='DC' then INFM  else 0 end,
CHEMEH=case when AID='DC' then HEMEH  else 0 end
INTO #DIAGIND2
FROM #DIAGIND

SELECT DISTINCT
A.MEMID
,A.MEDICAID_ID
,AID
,CATEGORY
INTO #RXCLAIMS
FROM ILLINOIS_REPORT_DETAILS.DBO.IL_FT_RX_CLAIMS A
JOIN [Illinois_Report_Details].[dbo].[IL_DIM_REF_GROUPER_MASTER] b
           on A.NDC = B.CODE
         and MODEL = 'MRX6.2.2R'
JOIN #MBRSHIP ON #MBRSHIP.MEMID = A.MEMID
WHERE DATEFILL BETWEEN '2015-07-01' AND '2016-06-30'
```

```
UNION

SELECT DISTINCT
MEMID
,medicaid_id
,AID
,CATEGORY
FROM [DC01DSSDBPC10\SQL10].[ERR_IL].[hds].[Pharmacy] a
JOIN [Illinois_Report_Details].[dbo].[IL_DIM_REF_GROUPER_MASTER] b
            on A.[NationalDrugCd] = B.CODE
          and MODEL = 'MRX6.2.2R'
JOIN #MBRSHIP ON #MBRSHIP.MEMID = A.[RecipientID]
WHERE [ServiceFromDt] BETWEEN '2015-07-01' AND '2016-06-30'
UNION
SELECT DISTINCT
MEMID
,medicaid_id
,AID
,CATEGORY
FROM [DC01DSSDBPC10\SQL10].[ERR_IL].hds.MMP_Pharmacy a
JOIN [Illinois_Report_Details].[dbo].[IL_DIM_REF_GROUPER_MASTER] b
            on A.[NationalDrugCd] = B.CODE
          and MODEL = 'MRX6.2.2R'
JOIN #MBRSHIP ON #MBRSHIP.MEMID = A.[RecipientID]
WHERE [ServiceFromDt] BETWEEN '2015-07-01' AND '2016-06-30'

SELECT memid,Aid, isnull([MRX1],0)MRX1, isnull([MRX2],0)MRX2,
      isnull([MRX3],0)MRX3, isnull([MRX4],0)MRX4,
 isnull([MRX5],0)MRX5, isnull([MRX6],0)MRX6, isnull([MRX7],0)MRX7,
      isnull([MRX8],0)MRX8, isnull([MRX9],0)MRX9,
 isnull([MRX10],0)MRX10, isnull([MRX11],0)MRX11,
      isnull([MRX12],0)MRX12, isnull([MRX13],0)MRX13,
      isnull([MRX14],0)MRX14,
  isnull([MRX15],0)MRX15
into  #Rx_Diagind
FROM
(SELECT distinct memid,Aid, CATEGORY FROM #RXCLAIMS) AS SourceTable
PIVOT(count(CATEGORY) FOR CATEGORY IN
      ([MRX1],[MRX2],[MRX3],[MRX4],[MRX5],[MRX6],[MRX7],[MRX8],[MRX9],
      [MRX10],
[MRX11],[MRX12],[MRX13],[MRX14],[MRX15] )  ) pvt

update #Rx_Diagind set mrx8=0, mrx7=0  where mrx9=1
update #Rx_Diagind set mrx7=0  where mrx8=1

SELECT DISTINCT A.*
,isnull([MRX1],0)MRX1, isnull([MRX2],0)MRX2, isnull([MRX3],0)MRX3,
      isnull([MRX4],0)MRX4,
 isnull([MRX5],0)MRX5, isnull([MRX6],0)MRX6, isnull([MRX7],0)MRX7,
      isnull([MRX8],0)MRX8, isnull([MRX9],0)MRX9,
 isnull([MRX10],0)MRX10, isnull([MRX11],0)MRX11,
      isnull([MRX12],0)MRX12, isnull([MRX13],0)MRX13,
      isnull([MRX14],0)MRX14,
  isnull([MRX15],0)MRX15
INTO #FINAL
FROM #DIAGIND2 A
```

```
LEFT JOIN #RX_DIAGIND B ON A.MEMID = B.MEMID

 update #FINAL set MRX1=0, MRX2=0 where CARVH >0 or CARM >0
 update #FINAL set CARL=0, CAREL=0, MRX2=0 where MRX1=1
 update #FINAL set MRX2= 0 where CARL>0 or CAREL>0
 update #FINAL set MRX3=0 where PSYH  >0 or PSYM  >0 or PSYML>0 or
      PSYL>0
 update #FINAL set MRX4=0 where DIA1H  >0 or DIA1M   >0 or DIA2M  >0
      or DIA2L>0
 update #FINAL set MRX5=0 where RENEH >0 or  RENVH >0
 update #FINAL set RENM=0, RENL=0 where MRX5=1
 update #FINAL set  MRX6=0 where HEMEH=1
 update #FINAL set HEMVH=0, HEMM=0, HEML=0 where MRX6=1
 update #FINAL set MRX9=0, MRX8=0, MRX7=0 where AIDSH >0 or INFH >0
 update #FINAL set HIVM=0 where MRX9=1
 update #FINAL set MRX8=0, MRX7=0 where HIVM=1
 update #FINAL set INFM=0, INFL=0 where MRX7   >0 or MRX8   >0 or
      MRX9>0
 update #FINAL set MRX10=0 where SKCM  >0 or SKCL    >0 or SKCVL >0
 update #FINAL set MRX11=0 where CANVH   >0 or CANH    >0 or CANM  >0
 update #FINAL set CANL=0 where MRX11=1
 update #FINAL set MRX12=0, MRX13=0, MRX14=0 where CNSH  >0 or CNSM >0
 update #FINAL set CNSL=0, MRX13=0, MRX14=0 where MRX12=1
 update #FINAL set MRX13=0, MRX14=0 where CNSL=1
 update #FINAL set MRX13=0 where MRX14=1
 update #FINAL set MRX15=0 where PULVH   >0 or PULH   >0 or PULM >0 or
      PULL >0
 update #FINAL set CARM=1 where AID='DC' and MRX1=1
 update #FINAL set HEMEH=1 where AID='DC' and MRX6=1
 update #FINAL set HIVM=0, INFM=0 where AID='DC' and MRX9=1
 update #FINAL set HIVM=1, INFM=0 where AID='DC' and ( MRX7 >0 or
      MRX8>0)

SELECT MEMID, AID, CATEGORY, Value
INTO #FINAL2
FROM (SELECT MEMID, AID,
      Intercept,a_under1,a_1_4,a_5_14m,a_5_14f,a_15_24m,a_15_24f,a_25_
      44m,a_25_44f,a_45_64m,a_45_64f,a_65,[CARVH],[CARM],[CARL],[CAREL
      ],[PSYH],[PSYM],[PSYML],[PSYL],[SKCM],[SKCL],[SKCVL],[CNSH],[CNS
      M],
[CNSL],[PULVH],[PULH],[PULM],[PULL],[GIH],[GIM],[GIL],[DIA1H],[DIA1M],
      [DIA2M],[DIA2L],[SKNH],[SKNL],[SKNVL],[RENEH],[RENVH],[RENM],
[RENL],[SUBL],[SUBVL],[CANVH],[CANH],[CANM],[CANL],[HLTRNS],[DDM],[DDL
      ],[GENEL],[METH],[METM],[METVL],[PRGCMP],[PRGINC],[EYEL],[EYEVL]
      ,[CERL],
[AIDSH],[INFH],[HIVM],[INFM],[INFL],[HEMEH],[HEMVH],[HEMM],[HEML],CCAR
      VH, CCARM,  CCNSH, CPULVH, CPULH, CGIH, CMETH, CHIVM, CINFM,
      CHEMEH,[MRX1],[MRX2],[MRX3],[MRX4],[MRX5],[MRX6],[MRX7],[MRX8],[
      MRX9],[MRX10],
[MRX11],[MRX12],[MRX13],[MRX14],[MRX15] FROM #FINAL) P
UNPIVOT
(VALUE FOR CATEGORY IN
      (Intercept,a_under1,a_1_4,a_5_14m,a_5_14f,a_15_24m,a_15_24f,a_25
      _44m,a_25_44f,a_45_64m,a_45_64f,a_65,[CARVH],[CARM],[CARL],[CARE
      L],[PSYH],[PSYM],[PSYML],[PSYL],[SKCM],[SKCL],[SKCVL],[CNSH],[CN
      SM],
```

```
[CNSL],[PULVH],[PULH],[PULM],[PULL],[GIH],[GIM],[GIL],[DIA1H],[DIA1M],
      [DIA2M],[DIA2L],[SKNH],[SKNL],[SKNVL],[RENEH],[RENVH],[RENM],
[RENL],[SUBL],[SUBVL],[CANVH],[CANH],[CANM],[CANL],[HLTRNS],[DDM],[DDL
      ],[GENEL],[METH],[METM],[METVL],[PRGCMP],[PRGINC],[EYEL],[EYEVL]
      ,[CERL],
[AIDSH],[INFH],[HIVM],[INFM],[INFL],[HEMEH],[HEMVH],[HEMM],[HEML],CCAR
      VH, CCARM,  CCNSH, CPULVH, CPULH, CGIH, CMETH, CHIVM, CINFM,
      CHEMEH,[MRX1],[MRX2],[MRX3],[MRX4],[MRX5],[MRX6],[MRX7],[MRX8],[
      MRX9],[MRX10],
[MRX11],[MRX12],[MRX13],[MRX14],[MRX15]) ) UNPVT
WHERE
VALUE = 1

DROP TABLE #IL_DIM_CDPS
SELECT DISTINCT
MEMID, #final2.AID, CATEGORY
, [CDPS_Label]
, ACUTE
,GETDATE() AS INSERT_DATE
INTO #IL_DIM_CDPS
FROM #FINAL2
LEFT JOIN ILLINOIS_REPORT_DETAILS.[dbo].[IL_DIM_REF_CDPS_WEIGHTS] B ON
      B.CDPS = #FINAL2.Category
        AND B.AID = CASE WHEN #FINAL2.AID IN ('DA','DC') THEN 'DADC'
      ELSE #FINAL2.AID END
        AND MODEL = 'PRO-RX'
ORDER BY 1

--Categories for regression testing
DROP TABLE #MRX
DROP TABLE #INTERCEPT
DROP TABLE #DEMOGRAPHIC
DROP TABLE #MEDICAL

SELECT DIStinct MEMID, SUM(ACUTE) MRX INTO #MRX from #IL_DIM_CDPS
WHERE category like '%MRX%'
GROUP BY MEMID

SELECT DIStinct MEMID, SUM(ACUTE)INTER INTO #INTERCEPT from
      #IL_DIM_CDPS
WHERE category like '%intercept%'
GROUP BY MEMID
SELECT DIStinct MEMID, SUM(ACUTE) DEMO INTO #DEMOGRAPHIC from
      #IL_DIM_CDPS
WHERE category like 'a_%'
GROUP BY MEMID

SELECT DIStinct MEMID, SUM(ACUTE) MEDI INTO #MEDICAL from #IL_DIM_CDPS
WHERE (category not like '%MRX%' and category not like '%intercept%'
      and category not like 'a_%')
GROUP BY MEMID

--creating 3 sets of data (chf , diabetes and breast cancer)
drop table #cdpsCAT
select memid, cdps_label
into #cdpsCAT
from(
```

```
SELECT DISTINCT
              MEMID, cdps_label, DENSE_RANK () OVER (PARTITION BY
      MEMID ORDER BY ACUTE DESC, CDPS_LABEL) AS RANKING
              FROM #IL_DIM_CDPS
 where (cdps_label like '%cancer%' OR cdps_label like
      '%Cardiovascular%' OR cdps_label like '%diabetes%')
              )ab
where ranking = 1

--Final Risk Score
DROP TABLE #Test_Data
SELECT DISTINCT A.MEMID
      ,[sex] GENDER
      ,[ETHINICITY]
      ,A.[AGE]
        ,DEMO
        ,INTER
        ,ISNULL (MEDI,0.0) MEDI
        ,ISNULL (MRX,0.0) MRX
        ,E.CDPS_SCORE STATE_SCORE
      ,C.[CDPS_NEW_RISK_SCORE] SCORE
      ,CASE WHEN D.[CDPS_LABEL] LIKE '%diabetes%' THEN 'DIABETES'
        WHEN D.[CDPS_LABEL] LIKE '%cardio%' THEN 'CHF'
        WHEN D.[CDPS_LABEL] LIKE '%cancer%' THEN 'BREAST CANCER'  END
      AS 'LABEL'
        ,CASE WHEN [SMOKER_6M]='Y' THEN 1 ELSE 0 END AS [SMOKER_6M]
      ,[IP_ADMIT_6M]
      ,[BED_DAYS_6M]
      ,[IP_BH_ADMIT_6M]
      ,[BED_DAYS_BH_6M]
      ,[ED_VISIT_6M]
      ,[PREVENTIVE_VISIT_6M]
      ,[HOSPITAL_PAID_AMT_6M]
      ,[HOSPITAL_ED_PAID_AMT_6M]
      ,[RX_PAID_AMT_6M]
INTO #Test_Data
FROM IL_DIM_MEMBER A
JOIN MEMBER B ON A.MEMID =B.MEMID
JOIN
      (SELECT DISTINCT
              MEMID
              , round(SUM(cast(ACUTE as real)),2) AS
      CDPS_NEW_RISK_SCORE
              FROM #IL_DIM_CDPS
              GROUP BY MEMID
      )C ON C.MEMID=B.MEMID
JOIN
      #cdpsCAT D on D.MEMID = B.MEMID
JOIN
      sfy2016_rs E ON A.MEMID = E.MEMID
JOIN
      MEMBER F ON A.MEMID = F.MEMID
LEFT JOIN
      #MRX ON A.MEMID = #MRX.MEMID
LEFT JOIN #INTERCEPT
       ON A.MEMID = #INTERCEPT.MEMID
LEFT JOIN #DEMOGRAPHIC
```

```
       ON A.MEMID = #DEMOGRAPHIC.MEMID
LEFT JOIN #MEDICAL
       ON A.MEMID = #MEDICAL.MEMID

--Preparing  data for FFDT test
DROP TABLE #DM_DATA
SELECT
A.AGE,
CASE
      WHEN GENDER='M' THEN 1 ELSE 0 END AS GENDER,
A.SCORE, A.SMOKER_6M SM,
CASE WHEN SCORE >5 THEN 3 --HIGH
       WHEN SCORE BETWEEN 2 and 5  THEN 2 --MEDIUM
       WHEN SCORE <2  THEN 1 --LOW
       END AS RISK_CAT,
CASE WHEN PROGRAM_RISK_LEVEL IS NULL THEN 0 ELSE 1 END CL_DM_FLAG,
B.ED_VISIT_6M ED_VI, B.[IP_ADMIT_6M] IP_AD ,READMISSION_6M_AUTH RE_AD
      , B.PREVENTIVE_VISIT_6M PCP_VI
INTO #DM_DATA
FROM RS_F A JOIN il_DIM_member B
ON A.MEMID=B.MEMID

--Add predicted DM logic to Risk Score Algorithm

Select
     CASE WHEN RISK_CAT=3 and (ED_VISIT_6M >= 2 OR ([IP_ADMIT_6M] >
      1)) and PREVENTIVE_VISIT_6M >1 THEN 'Tier I'
      WHEN RISK_CAT=2 and (ED_VISIT_6M < 2 OR ([IP_ADMIT_6M] < 2)) and
      PREVENTIVE_VISIT_6M >1 THEN 'Tier II'
      ELSE 'Tier III' END AS DM_LEVEL,
        A.*
FROM #Test_Data A
JOIN
      #DM_DATA B ON A.MEMID =B.MEMID
```

7. Bibliography

[1] Chronic Disease Prevention and Health Promotion Available: https://www.cdc.gov/chronicdisease/overview/index.htm#ref1

[2] BriSimi, S. T., Xu , T., Wang, T., Dai, W., Adams, G. W., & PaSCh , Ch. I., *Fellow IEEE* , 2017. Predicting Chronic Disease Hospitalizations from Electronic Health Records: An Interpretable Classification Approach. *Proceedings of the IEEE*, 1-18, doi:10.1109/JPROC.2017.2789319.

[3] Liang J. B., Ma M. X., Wu J. J., Yan D. J., 2017. Design and construction of mobile chronic diseases management system, *2017 2nd IEEE international conference on computational intelligence and applications*, 518- 522.

[4] Hoof C. V., 2017. The virtual personal health coach: technology and data analytics join forces to disrupt preventive health, 987-1-5090-6707-7/17/2017 IEEE, 233-233, doi: 10.1109/IWASI.2017.7974258.

[5] Disease and Health Risk Management. Available: https://www.harvardpilgrim.org/portal/page?_pageid=253,41752&_dad=portal&_schema11=PORTAL

[6] Li L., Bagheri S., Goote H., Hasan A., Hazard G., 2013, Risk Adjustment of Patient Expenditures: A Big Data Analytics Approach, 2013 IEEE International Conference on Big Data, 12-14.

[7] American academy of actuaries, 2010. Risk Assessment and Risk Adjustment, *Issue brief*, May 2010, 1-8.

[8] Stranieri A., Yatsko A., Venkatraman S. & Jelinek F. H., 2018. Data Analytics to Select Markers and Cut- off Values for Clinical Scoring. *In Proceedings of Australasian Computer Science Week 2018, Brisbane, QLD, Australia, January 2018 (ACSW 2018)*, 1-6, doi: 10.1145/3167918.3167931.

[9] 2018 ICD-10-CM Codes. Available: http://www.icd10data.com/ICD10CM/Codes

[10] Liu X., Li B. X., Motiwalla L., Li W., Zheng H., & Franklin D. P., 2016. Preserving patient privacy when sharing same-disease data. J. Data Information Quality, Vol. 7, No. 4, Article 17, September 2016, 1- 14, doi: http://dx.doi.org/10.1145/2956554

[11] Pazhaniraja N., Paul P.V., Priyadharshini C., Maris M. H., Leticianathali A.. 2017. A Survey on Various Prediction model for Chronic Obstructive Pulmonary Disease (COPD), *2017 International Conference on Computation of Power, Energy, Information and Communication (ICCPEIC)*, 269-272.

[12] Fung G. Y., Dy J. G., & Rosales, R. 2010. Medical coding classification by leveraging inter-code relationships. *In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM, New York, 193–202.

[13] Hussein S. A., Omar M. W., Li X., Ati M., 2012. Efficient Chronic Disease Diagnosis Prediction and Recommendation System, *2012 IEEE EMBS International Conference on Biomedical Engineering and Sciences I Langkawi I 17th - 19th December 2012*, 209-214.

[14] Hung Y. C., Chen C. W., Lai T. P., Lin H. C., & Lee C. C., 2017. Comparing Deep Neural Network and Other Machine Learning Algorithms for Stroke Prediction in a Large-Scale Population-Based Electronic Medical Claims Database, Engineering in Medicine and Biology Society (EMBC*), 2017 39th Annual International Conference of the IEEE*, 3110-3113.

[15] Guruvayur R. S., Dr. Suchithra R, 2017. A detailed study on machine learning techniques for data mining, *International Conference on Trends in Electronics and Informatics ICEI 2017*, 1187-1192.

[16] Polaraju K., Prasad D.D., 2017. Prediction of Heart Disease using Multiple Linear Regression Model, 2017 IJEDR, Volume 5, Issue 4, ISSN: 2321-9939, 1419-1425.

[17] Ordonez C.,2006. Comparing Association Rules and Decision Trees for Disease Prediction, ACM, 2006, 1-8, doi: http://doi.acm.org/10.1145/1183568.1183573.

[18] Roy S., Mondal S., Asif Ekbal A., Desarkar S. M., 2016. CRDT: Correlation Ratio Based Decision Tree Model for Healthcare Data Mining, *2016 IEEE 16th International Conference on Bioinformatics and Bioengineering*, 36-46, doi: 10.1109/BIBE.2016.21.

[19] Katsikopoulos V. K., Member, IEEE, and Fasolo B., 2006. New Tools for Decision Analysts, *IEEE Transactions on Systems, Man, and Cybernetics, - Part A, Systems and Humans*, Vol. 36, No. 5, 960-967.

[20] Silva F. de .M., Isaksen A., Togelius J., Nealen A., 2016. Generating Heuristics for Novice Players, *Submission for IEEE Computational Intelligence in Games 2016*, 1-9.

[21] Mannino M., Fredrickson J., Kashani B. F., Linck I., & Alqurashi R, 2017. Development and evaluation of a similarity measure for medical event sequences. ACM

Trans. Manage. Inf. Syst. Vol. 8, No. 2–3, Article 8, June 2017, 1-26, doi: http://dx.doi.org/10.1145/3070684

[22] Yadav P., Steinbach M., Kumar V., & Simon G., 2018. Mining Electronic Health Records (EHRs): A Survey. ACM Comput. Surv. Vol. 50, No. 6, Article 85 (January 2018), 1-40, doi: https://doi.org/10.1145/3127881

[23] Wang F., Zhang P., Qian B., Wang X., Davidson I., 2014. Clinical Risk Prediction with Multilinear Sparse Logistic Regression, KDD'14, August 24–27, 2014, New York, New York, USA. Copyright © 2014 ACM 978-1-4503-2956-9/14/08, 145-156, doi: http://dx.doi.org/10.1145/2623330.2623754.

[24] Heart Disease Fact Sheet, Division for Heart Disease and Stroke Prevention. Available: https://www.cdc.gov/dhdsp/data_statistics/fact_sheets/fs_heart_disease.htm

[25] Khan U. M., Choi P. J., Shin H. & Kim M., 2008. Predicting Breast Cancer Survivability Using Fuzzy Decision Trees for Personalized Healthcare, *30th Annual International IEEE EMBS Conference Vancouver, British Columbia, Canada, August 20-24, 2008*, 5148-5153.

[26] Abreu H. P., Santos M. S., Abreu H. M., Andrade B., & Silva C. D., 2016. Predicting breast cancer recurrence using machine learning techniques: A systematic review. ACM Comput. Surv. Vol. 49, No. 3, Article 52 (October 2016), 1-40, doi: http://dx.doi.org/10.1145/2988544.

[27] American Cancer Society, Breast Cancer Facts & Figures 2017-2018. Available: https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/breast-cancer-facts-and-figures/breast-cancer-facts-and-figures-2017-2018.pdf

[28] National Center for Health Statistics. Available: https://www.cdc.gov/nchs/fastats/diabetes.htm

[29] National Diabetes Statistics Report, 2017, Estimates of Diabetes and Its Burden in the United States. Available: https://www.cdc.gov/diabetes/pdfs/data/statistics/national-diabetes-statistics-report.pdf

[30] Kronick R., Ph.D, Gilmer P. T., Ph.D, Dreyfus T., M.C.P, Ganiats G. T., M.D, 2002. CDPS-Medicare: The Chronic Illness and Disability Payment System Modified to Predict Expenditures for Medicare Beneficiaries, *final report to CMS June 24, 2002* .1-147.

[31] Kelarev A.V., Stranieri A. Yearwood J.L, Jelinek H.F., 2012. Empirical Study of Decision Trees and Ensemble Classifiers for Monitoring of Diabetes Patients in Pervasive Healthcare, 2012 IEEE, 441- 446, doi: 10.1109/NBiS.2012.20.

[32] Kumari D., Dr. S. Seema, 2016. Predictive Analytics to Prevent and Control Chronic Diseases, *2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), 978-1-5090-2399-8/16, 2016 IEEE*, 381-386.

[33] Eftekhari S., Yaraghi N., Singh R., Gopal D. R., & Ramesh R., 2017. Do health information exchanges deter repetition of medical services? ACM Trans. Manage. Inf. Syst., vol. 8, No. 1, Article 2, April 2017, 27-37, doi: http://dx.doi.org/10.1145/3057272.

[34] Chini F, Pezzotti P., Orzella L., Borgia P. & Guasticchi G., 2011. Can we use the pharmacy data to estimate the prevalence of chronic conditions? a comparison of multiple data sources, 1-8.

[35] Gorina Y. & A. Kramarow A. E., Identifying Chronic Conditions in Medicare Claims Data: Evaluating the Chronic Condition Data Warehouse Algorithm, 2011. Health Research and Educational, 1610- 1627, doi: 10.1111/j.1475-6773.2011.01277.x

[36] Juhnke C., Bethge S. & Mühlbacher C. A., 2016. A Review on Methods of Risk Adjustment and their Use in Integrated Healthcare Systems, 1–18, doi: http://dx.doi.org/10.5334/ ijic.2500

[37] Fehring K. T., MD, AAHKS Risk Adjustment Initiative: Why Is It Important? 2016. *AAHKS Symposium: Patient Reported Outcome Measures: This is your New Reality*, 1148- 1150, doi: http://dx.doi.org/10.1016/j.arth.2016.02.083

[38] Geruso M. & McGuire G. T., 2016. Tradeoffs in the design of health plan payment systems: Fit, power and balance, 1-19, doi: http://dx.doi.org/10.1016/j.jhealeco.2016.01.007

[39] Song S., Warren J., Patricia R., 2014. Developing high risk clusters for chronic disease events with classification association rule mining, *Proceedings of the Seventh Australasian Workshop on Health Informatics and Knowledge Management (HIKM 2014), Auckland, New Zealand*, 69-78.

[40] Wickramasinghe K. L., Alahakoon D., Georgeff M., Schattner P., De Silva D., Alahakoon O., Ada ji1 A., Kay Jones K., Piterman L., AM, 2011. Chronic Disease Management: a Business Intelligence Perspective, 1-8.

[41] Sunyaev A. & Chornyi D., 2012. Supporting chronic disease care quality: Design and implementation of a health service and its integration with electronic health records. ACM J. Data Inf. Qual. Vol 3, No. 2, Article 3 (May 2012), 1-21, doi = 10.1145/2184442.2184443 http://doi.acm.org/10.1145/2184442.2184443