Capitation Cost Prediction and

Visualization

TCSS 702 MS Project Report

**Seyed Mehdi Jazayeri**

Master of Computer Science and Systems

Institute of Technology

University of Washington, Tacoma

**Community Members**:

Dr. Ankur Teredesai, Chair

Dr. Shanu Sushmita

Spring - Summer 2016

Table of Contents:

[Abstract:](#h.5ae4jx3eyy26)

[1. Introduction](#h.2ijd1g1gmp4g)

[1.1. Capitation Payment](#h.2ufrlpjslr05)

[1.1.1. Risk Adjustment](#h.ltbhu47enu7z)

[2. Methodology:](#h.hjlxboyuqwvo)

[2.1. Dataset](#h.5qrpyxohngl3)

[2.2. Data Cleaning and Preprocessing](#h.a74ie3lc2mcq)

[2.3. Method](#h.xozu1ht59tjd)

[2.3.1. DCG/HCC Models](#h.x77zrxmgiafc)

[2.3.2. Diagnostic Groups (DxGroups)](#h.ehm35k3p0yhr)

[2.3.3. Condition Categories](#h.21i0uvlreqec)

[2.3.4. CC Hierarchies](#h.2m7v17dpluil)

[2.4. Evaluation](#h.i22eg75h9ci3)

[3. System Architecture:](#h.zfpkorczx8zs)

[3.1 Data Layer](#h.ze9id0wh0l9z)

[3.2. Model Bank](#h.44wtwkbmmr07)

[3.2.1 Demographic Model](#h.bk4se1m0nbxy)

[3.2.2. Linear Model](#h.lvmusfc6z63n)

[3.2.3. Decision Tree Model](#h.62mnk08hne6l)

[3.3 Plotter](#h.t5z9xjf2qaqy)

[3.4 Web Application](#h.xcixc1ul6ubi)

[3.4.1. User Interface (UI)](#h.bzab39iab5je)

[3.4.2. Server](#h.dypw3fkobq6y)

[4. Findings of the project](#h.9l11b6o9dwdp)

[5. Challenges and Obstacles](#h.b401cekmj8ln)

[6. Conclusion](#h.bbofaq5ppwqd)

[References](#h.n3mx6jynz8md)

# Abstract:

The rising cost of healthcare is one of the world’s most important problems. Since the 1980s, there has been research on the predictive modeling of medical costs based on health insurance claims as a way to control cost growth. In addition, there is an increasing interest on the part of healthcare payers and providers to develop statistical models that can accurately predict future healthcare expenditures. The change from fee-for-service (FFS) payment systems to capitation payment is considered as a viable solution to this problem. The proposed payment system is based on risk-adjustment for a population of patients. Having not precise risk-adjustment system, this system would have even more adverse effect on healthcare expenses. This project is an implementation of DCG/HCC capitation payment method which utilize demographic and diagnostic information of patients to construct a cost prediction model. The underlying approach of DCG/HCC systems is to form homogeneous subpopulation of patients based on diagnostic information, then, build a predictive model. In addition, it provides tools for visualizing the population’s attributes to facilitate data analysis.

# 

# 1. Introduction

Healthcare spending in the U.S. today is not sustainable. Average healthcare spending per capita in the U.S. exceeds $8,000 [1]. One of the proposed solution to control this increasing expenditures is to reconsider methods of payment. The current predominant fee-for-service (FFS) based systems has a significant impact on the increased overheads. In this model, providers are paid for each service rendered during the treatment. Due to the fact that this payment model is dependent on the quantity of service rather than quality of service, physicians and providers have the incentive to provide more treatments than what it is required. Furthermore, in FFS system, providers bear a negligible financial risk (as they are paid on the basis of services provided) while insurer deal with the high risk of losing money (payments to providers can exceed the amount of premiums collected). Therefore, reforms in current system of payment is high on the political agenda. The purpose of these reforms is to make resource allocation in healthcare more efficient. One of the proposed systems of payment to tackle this issue is capitation payment system.

## 1.1. Capitation Payment

The word “capitation” is derived from the term “per capita” which means per person. In this system, physicians and providers are paid a set amount per member per month (PMPM) whether the person seeks care or not. In this definition, member means enrollee in some managed care plan, which is usually a health maintenance organization (HMO). For example, a primary care physician may receive $20 PMPM for providing the healthcare needs of 100 enrollee of a regional HMO. Under this contract, the physician receives 20 \* 100 \* 12 = $24,000 in total capitation payment over a year and this amount must cover all the primary care services offered to the patient population specified in the contract.

Under capitation, risk sharing occurs between two parties; provider and insurer. Providers bear the risk that the cost of providing services might exceed the capitation payments. Insurers bear the risk that provider costs can increase when contracts are renewed.

In addition, capitation reverse the actions providers must take to achieve financial success. Under FFS, keys to success are to work harder and increase volume in order to increase profits. So, the primary task of managers is to maximize utilization and reimbursements. On the other hand, under capitation, keys to profitability are to work efficient and decrease volume. In general, capitation motivates provider to provide only necessary services, and to provide those services in the lowest possible cost. Here are some potential benefits associated with capitation [2]:

* Providers receive a fixed payment regardless of whether services are actually rendered.
* Capitation revenues are predictable and timely, and thus are less risky than conventional payment methodologies.
* Capitation payments are received before services are rendered, so, in effect, payers are extending credit to providers rather than vice versa, as under conventional reimbursement.
* Capitation supports national healthcare goals; primarily increased emphasis on cost control as well as wellness and prevention.
* Capitation may ease the reimbursement paperwork burden, and hence reduce expenditures on administrative costs.
* Capitation aligns the economic interests of physicians and hospitals because risk-sharing systems are typically established that allow all providers in a capitated system to benefit from reducing costs.
* Similarly, capitation encourages utilization of lower-cost treatments, such as outpatient surgery and home health care, as opposed to higher-cost inpatient alternatives.

### 1.1.1. Risk Adjustment

The capitation payment should be adjusted to the healthcare needs of members. The payment per enrollee is dependent of the risk category to which the person belongs. Thus, the success of capitation model hinges on an accurate risk-adjustment system [3]. In order to precisely calculate capitation cost of enrollees, an accurate risk-adjustment system would divide enrollees into homogeneous risk-groups such that members of each group are similar in costs. When the risk-groups are rather heterogeneous the capitation system has two disadvantages. First, risk-groups that are too heterogeneous may result in an unfair distribution of payments. A provider with relatively unhealthy members per risk group will be underpaid. The second disadvantage is that cream skimming may be very advantageous to providers (cream skimming is selection of less costly enrollees). The development of such a system is a crucial technical problem and is intended to provide efficiency and fairness for capitation model.

In the literature several risk-adjustment (classification) systems have been suggested for capturing the future costs of inpatient morbidity. Ash [4][5] constructed Diagnostic Cost Groups (DCGs). Ellis [6] and Ash [7] further refined the Diagnostic Cost Group model. Ash and several coworkers developed the DCG model, using both clinical and economic criteria. Pope introduced CMS hierarchical condition categories (HCC) model in 2004 to adjust Medicare capitation payments to private health care plans for the health expenditure risk of their enrollees.

# 2. Methodology:

## 2.1. Dataset

The California Office of Statewide Health Planning and Development (OSHPD) provides public data sets of inpatient data collected from California-licensed hospitals in California. The data set consists of records of all inpatient discharged from a California-licensed hospital. Licensed hospitals include general acute care, acute psychiatric, chemical dependency recovery, and psychiatric health facilities. We have data available from year 2009 to 2013 and for the purpose of this project I used 2012 data to predict 2013 costs.

Each record consist of 130 attributes including demographic attributes of patients like age, gender, ethnicity, race, etc. In addition, it contains information like admission date, length of stay, type of care and so on. The most important attribute, Charge, is defined as total costs, which includes all costs for services rendered during the length of stay for patient care at the facility, based on the hospital’s full established rates. Costs include, but are not limited to, daily hospital services, ancillary services and any patient care services. Hospital-based physician fees are excluded. Prepayments (e.g., deposits and prepaid admissions) are not deducted from the total Cost. Following diagrams are representing important information of OSHPD train dataset that are used in our capitation payment modeling.

|  |  |
| --- | --- |
| TrainAgeDistribution.png | TrainCostDistribution.png |
| a)Age Distribution | b) Cost Distribution |
| TrainGenderDistribution.png | TrainTop10HCC.png |
| c) Gender Distribution | d) Top 10 HCC Distribution |

Figure 1. Dataset population and cost statistics

### 2.2. Data Cleaning and Preprocessing

Before the dataset can be used for modeling, it must be processed and prepared. These are the undertaken steps in data cleaning and preprocessing:

1. Removing records with invalid values: In OSHPD dataset there are records with null, zero or N/A values. They can be categorized as follow:
2. Empty, N/A or 0 cost values.
3. Negative age values.
4. Values other than 0 and 1 for gender.
5. Invalid ICD-9-CM codes.
6. Adjusting attribute types: Type adjustment is one of the most important part of this model without which the final result cannot be trusted. For example, Age is a numeric variable that must be converted into a factor variable. After this conversion, age is a factor variable with the following levels:

[0-5], [6-12], [13-17], [18-24], [25-34], [35-44], [45-54], [55-64], [65-69], [70-74], [75-79], [80-84], [85-89] and, [90+].

1. Mapping Diagnostic Codes To HCC Codes: The next step is to create a mapping from diagnostic codes(ICD-9-CM) to HCC. To achieve this goal one must first assign each diagnostic to its related CC. Then hierarchies are imposed to put CCs into hierarchies (Further explanations are provided under session 2.3). The final result of this mapping is an assignment of ICD-9-CM codes to HCC groups. From this mapping file, a csv file is generated as a lookup table to find HCC codes of patients.

## 2.3. Method

Historically, capitation systems were based on merely demographic information of patients. Not being able to provide homogeneous subpopulations, these demographic risk-adjusters failed to accurately predict the next year’s expenditures. This disadvantage has undesirable consequences for cost management systems. On the other hand, it gives private sectors the incentive to accept coverage of healthier enrollees, for whom the predicted costs are higher than their actual healthcare needs [8]. DCG/HCC models are invented to resolve the problem of heterogeneous subpopulations.

### 2.3.1. DCG/HCC Models

Next generations of risk-adjuster, DCG models, are designed to utilize diagnostic information (ICD-9-CM) of enrollees in addition to their demographic information. The idea is to generate subpopulations based on diagnostic information so that patients in same subpopulation are cost similar [9]. In these models predictions are based on diagnostic codes rather than procedure or episode of care. DCG models express a person’s health from his or her CCs and estimate expected costs.

### 2.3.2. Diagnostic Groups (DxGroups)

More than 15000 ICD-9-CM codes are too fine to be directly used as a payment classification system. Thus, they are grouped together to form 543 categories, called DxGroups which are basic building blocks of DCG/HCC models. Each DxGroups has a two-level numerical label and a short, clinically informative text name. All DxGroups with the same “whole number” stem are clinically related. Each ICD-9-CM code map to a unique DxGroup [9].

### 2.3.3. Condition Categories

Medically related DxGroups with similar expected costs are further clustered into a Condition Category (CC). Then, the resulting 118 diagnostic-based CCs are used in modeling. Furthermore, logical inconsistencies in diagnostic coding that can be identified by comparing with age and sex are eliminated from categories. As an example, a diagnosis of uterine disorder in a male is dropped from the model. Also, some CCs are entirely excluded from the model because of their weak effect on the next year’s costs.

### 

### 2.3.4. CC Hierarchies

A payment model should not be sensitive to every diagnostic code recorded because this will result in unstable estimates of the relative risk of populations. Hierarchies are used to constrain CC assignment so that a person classified into a CC is not also classified into a lower ranked CC in the same hierarchy. In another word, hierarchies reduce the sensitivity of predicted payments to the coding of less serious manifestations of the same condition [10]. For a female with both cancer and diabetes, hierarchies are used to retain only the “worst” evidence of each disease, but both cancer and diabetes CCs are used in predicting her costs. Moreover, the hierarchies prevents rewarding plans that capture as many codes as can be legitimately defended in an audit. The CC hierarchies not only capture chronic and serious acute manifestations of particular disease processes, but also capture their seriousness in terms of expected costs [9].

Following principles are guiding this classification system [10]:

**Principle 1)** Condition categories should be clinically meaningful. Each condition category is a set of ICD-9-CM codes. These codes should all relate to a reasonably well-specified disease or medical condition that defines the category.

**Principle 2)** CCs should predict medical expenditures. Diagnoses in the same HCC should be reasonably homogeneous with respect to their effect on both current (this year’s) and future (next year’s) costs.

**Principle 3)** **CCs that will affect payments should have adequate sample sizes to permit accurate and stable estimates of expenditures**.

**Principle 4)** In creating an individual’s clinical profile, hierarchies should be used to characterize the person’s illness level within each disease process, while the effects of unrelated disease processes accumulate. Related conditions should be treated hierarchically, with more severe manifestations of a condition dominating less serious ones.

**Principle 5)** The diagnostic classification should encourage specific coding. Vague diagnostic codes should be grouped with less severe and lower-paying diagnostic categories to provide incentives for more specific diagnostic coding.

**Principle 6)** The diagnostic classification should not reward coding proliferation. The classification should not measure greater disease burden simply because more ICD-9-CM codes are present. Hence, neither the number of times that a particular code appears, nor the presence of additional, closely related codes that indicate the same condition should increase predicted costs.

**Principle 7)** Providers should not be penalized for recording additional diagnoses. This principle has two consequences for modeling: (1) no condition category should carry a negative payment weight, and (2) a condition that is higher-ranked in a disease hierarchy should have at least as large a payment weight as lower-ranked conditions in the same hierarchy.

**Principle 8)** The classification system should be internally consistent. If diagnostic category A is higher-ranked than category B in a disease hierarchy, and category B is higher-ranked than category C, then category A should be higher- ranked than category C. This ensures that the assignment of diagnostic categories is independent of the order in which hierarchical exclusion rules are applied.

**Principle 9)** The diagnostic classification should assign all ICD-9-CM codes. Since each diagnostic code potentially contains relevant clinical information, the classification should categorize all ICD-9-CM codes.

**Principle 10)** Discretionary condition categories should be excluded from payment models. Diagnoses that are particularly subject to intentional or unintentional discretionary coding variation or inappropriate coding by health plans/providers, or that are not clinically or empirically credible as cost predictors, should not increase cost predictions. Excluding these diagnoses reduces the sensitivity of the model to coding variation, coding proliferation, gaming, and upcoding.

When patients are assigned into their relevant HCCs, the data is ready for the next step; training the predictive model. Three modeling techniques are used and evaluated. In previous studies the main technique employed was a linear model which is using demographic and HCC attributes. However, we took one step further to train and test a Decision Tree model as well. Furthermore, the old, demographic model is implemented to serve as a baseline model.



Figure 2. Hierarchical Condition Categories Aggregations of ICD-9-CM Codes

## 2.4. Evaluation

Because implementing a risk-adjustment model has serious consequences, we must understand how well the models work. The one universally reported, single number summary performance measure for risk-adjustment payment models is the R2 [11] (or the proportion of variance in costs that the model explains) [12]. In our evaluation, R2 for each CC hierarchy (HCC) is calculated. The closer R2 to 1, the closer the predicted cost around the mean hence, the better the prediction. However, for some HCCs the R2 is negative. The main cause of negative R2 values is the lack of enough data points in the related HCCs (inconsistency with principle 10). In evaluation of the model, the traditional demographic model is utilized as the baseline. Table 1 shows the R2 evaluation of top 10 populated HCCs for three different models.

|  |  |  |  |
| --- | --- | --- | --- |
| **HCC** | **R2\_Demogrpahic** | **R2\_LinearHCC** | **R2\_DecisionTreeHCC** |
| 2 | -0.075913764 | 0.004633241 | 0.007750202 |
| 54 | -0.568518562 | 0.003812596 | 0.013938131 |
| 55 | -1.356974402 | 0.015745812 | 0.039803996 |
| 79 | -0.114383689 | -0.002708267 | -0.002252807 |
| 80 | -0.005801435 | 0.003138198 | 0.001718534 |
| 81 | -0.097383211 | 0.011504488 | 0.016533157 |
| 92 | -0.105059262 | -0.0093416 | -0.007246929 |
| 96 | -0.01049921 | -0.000294691 | -0.001127068 |
| 108 | -0.24715498 | -0.018739007 | -0.001489255 |
| 164 | -0.053382813 | 0.002310203 | 0.002298904 |

Table 1. R2 of 10 most populated HCC groups resulted by three models

# 

# 

# 3. System Architecture:

The general architecture of the system can be divided into four components; Data Layer, Model Bank, Plotter and, Web Application. Figure 1 shows how these components are connected.



Figure 3.Abstract System Architecture

## 3.1 Data Layer

This component is responsible for reading and preparing datasets. The main function in this module is *ReadData* which will perform all data cleaning and preparation processes described in session 2.2. Two datasets (train and test sets) are passed to this function before being used in any other component. Figure 2, illustrates the flow of data in this component.



Figure 4. Flow of data in Data Layer

## 3.2. Model Bank

Model Bank is the machine learning core of the system. Here, predictive modeling, prediction and evaluation are carried out. There are three models that user can select; Demographic Model, Linear Model and, Decision Three Model. Each of them has its own internal variable for predicted costs and R2 evaluations inside the component. These values are used by different components, specially, the Plotter. When a user uploads a new train set, this component is reset and all previous values will be deleted.

In addition to the above mentioned models, a Support Vector Machine (SVM) model was initially considered for the Model Bank. However, it is removed from models due to its poor performance on the current datasets.

### 3.2.1 Demographic Model

Demographic model is a linear model based on demographic information of patients (i.e age and gender). This model is added for performance evaluation of the other diagnostic based models. There are three internal variables inside this component associated with this model; *demographic.model, demographic.res and demographic.eval*. They retain model, predicted costs and, R2 of the demographic model respectively.



Figure 5. Model Bank Structure

### 3.2.2. Linear Model

Linear Model is the recommended model for DCG/HCC based capitation cost prediction. This model regress cost on age, gender, and HCC (cost ~ age + gender + hcc). In general, this model explains how changes in one feature (having the rest fixed) would affect cost. Three variables in *Model Bank* are associated to this model; *linear.model, linear.res and, linear.eval*. They hold values of model, predicted costs and, R2 of linear model respectively.

### 3.2.3. Decision Tree Model

A Decision Tree model is also included in the Model Bank to capture probable non-linear characteristics of the data [13]. Interestingly, for certain HCCs a Decision Tree model outperforms the Linear model. A reliable, already existing CRAN recommended package, *rpart,* is used to train the decision tree model.

Since overfitting has always been one of the critical issues of tree based models, a control parameter is used to constrain unnecessary growth of the three. This parameter is defined as follow:

**rpart.control(minsplit=2, minbucket=1, cp=0.0001)**

Once again, three variables in the *Model Bank* are associated to this model; *dtree.model, dtree.res and, dtree.eval*. They hold values of model, predicted costs and, R2 of decision tree model respectively.

## 3.3 Plotter

*Plotter* is the graphic core of the system and responsible for generating all the plots shown to the end user on the web application. This component, features general purpose functions to render plots for distribution of dataset’s features according to their types. Moreover, there are functions to plot distribution comparison of two variables from two distinct datasets A.K.A first year and second year (or, year1 and year2). I leveraged one of the best plotting CRAN recommended packages for creating static charts, *ggplot2.* This package is a data visualization package for *R* and is an implementation of *Leland Wilkinson's Grammar of Graphics;* a general scheme for data visualization which breaks up graphs into semantic components such as scales and layers. *ggplot2* can serve as a replacement for the base graphics in R and contains a number of defaults for web and print display of common scales. It is licensed under *GNU GPL v2*.

In general, there are three types of function in this component:

First, *plotDistribution.*[type]functions output distribution plot of a variable with specific type. These functions are used to generate plots for Age, Gender and, Cost distributions. While distribution of factor variables better understood with histograms, numeric variables can be understood with both a histogram or a density plot.

Second, *plotDistribution.*[type1]*\_*[type2] functions, render plots for distribution of a variable of type1 over a variable of type2. These functions are used in plotting the distribution of Cost over Age, Cost over Gender and, Cost over HCC. After examining different chart types, pie chart is selected as the best candidate for the purpose of these functions.

In addition to the above functions, *plotTopNDistribution* functions, are added to sketch effective plots for factor variables with more than 20 categories. As an example, HCC with 177 different categories was challenging to be plotted. Thus, top N values are selected and all the rest are labeled as “Others”. These functions output distribution of Cost over Age and Cost over HCC as well as distribution of HCC. Below are the steps taken for Distribution of numeric over factor:

**Step 1**: create a dataframe of the numeric variable and the factor variable

**Step 2**: calculate total value of numeric variable for each factor level

(aggregate(num ~ fact, FUN = sum).

**Step 3**: Sort the dataframe ascendingly by total value of numeric variable.

**Step 4**: Select top N rows.

To plot top N factor distribution plots, similar steps are taken. However, instead of calculating total value, frequency of each factor level is calculated.

Third, *plotComparision*.[type] functions, generate comparison charts for specific variables of two datasets. When it comes to numeric comparison, these functions have the option to create plot for a given range of values. Comparison distribution of Age, Gender, HCC and Cost between Year 1 and Year 2 are products of these functions.

## 3.4 Web Application

Shiny is framework based on R programming language. This framework is used to implement the web application of this project. It is an open source package that provides an elegant and powerful web framework for building web applications. Typically, a Shiny App follows the principle of “separation of concerns” by providing two separate files, *ui.R* and *server.R*. As their names suggest, user interface components are implemented in the first file and implementation details are defined in the second file. Following sections describe how capitation payment system is implemented in Shiny ecosystem.

### 3.4.1. User Interface (UI)

The whole user interface is implemented in a single file called *ui.R* which describes arrangement of different objects of the web UI. To access these objects, a unique ID is assigned to each of them. This file is logically separated into four divisions each of them correspond to a web page. To easily access web pages, their links are placed on the left menu.

The first division, FILE UPLOAD, consists of UI components in Upload Page. Second division, DATA ANALYSIS, defines where and how the distribution plots are going to be shown. Population distribution plots and Cost distribution plots are separated using a tab menu. Third division, contains arrangements of components of data comparison page. Last but not least, COST PREDICTION, contains elements of the final page.

### 3.4.2. Server

*server.R* holds the server side code of the web application. Main task of the server is to receive input data, perform operations on the data and connect different components together to provide responses to a user’s requests. To add more readability to the code, it is logically separated into four part. Each part is responsible for operations performed on pages mentioned in 3.4.1. The server calls functions in *Data Layer, Plotter* and *Model Bank* and produce contents of the application.



Figure 6. Shiny Web App Components

# 4. Findings of the project

One of the interesting findings of this project was that using more complex machine learning techniques like Tree based models, would yield more accurate results than simple linear model. However, for some HCC groups, a linear model outperforms decision tree model. Initially, I hoped that a Support Vector Machine (SVM) would generate even more accurate predictions than decision tree model. Nevertheless, svm produced unsatisfactory results compared to other models. Thus, it is removed from the final version of model bank. The lesson I learned from it is that complexity of a model doesn’t necessarily guarantee more accurate predictions.

Furthermore, I learned that there is a correlation between the number of patients in each HCC and the accuracy of cost prediction of that HCC. In that the more data in a HCC, the more accurate the predicted cost of the HCC.

Finally, observing R2 evaluations of two models (linear and decision tree), triggered the notion of using different modeling techniques for different HCCs. As of now, a single technique is used to make prediction for all HCCs. As it can be seen in Table 1, cost prediction of some HCCs are more accurate using decision tree while prediction of others are more accurate using just a linear model.

# 5. Challenges and Obstacles

In this project I learned a lot from the challenges I faced. At the beginning, as a newcomer, I had to learn numerous domain terminologies that I had never heard of before. The next step was to find related publications. Capitation systems are fairly new and still under study. Therefore, finding recent and reliable materials to understand the system was not trivial.

Furthermore, creating diagnostic hierarchies needs clinical domain knowledge which I do not posses. After a week of research I found a file in Medicare & Medicaid Services website which contains mapping of ICD-9-CM to DxGroups and HCCs. Even Though I had access to the mapping file, development of an efficient lookup table to assign HCCs to patients was another challenge. At the end I improved lookup operation from O(n) to O(1) which was an excellent achievement.

Another challenge was the development of web application. I was searching for R frameworks to develop the web app or JavaScript libraries that are compatible with R. Finally, I selected ShinyR framework for it is completely based on R and I had the chance to learn a new technology.

During the development of web application, I realized how complicated visualization techniques could be. I started to generate charts using base graphic functions of *R*. However, the outcome was not satisfactory enough. Additionally, plotting HCC distribution was a problem since there are 177 of them. Consequently, after studying numerous tutorials, I switched the plotting tools from default R package to *ggplot2*. Also, I decided to plot top 10 important HCC and aggregate of the rest(“Others”) to produce meaningful plots related to HCC.

Evaluation of the model was another challenge. When I first calculated R2, I ended up having negative values for almost all HCC groups. Then I realized, I was using a sample dataset and the dataset was not large enough to produce desired results. Still, for some less populated HCC groups the R2 is negative and the reason is the lack of enough data for those groups. It is good to mention that this system words best with “Claim Dataset” but our dataset, OSHPD, is a “Clinical Dataset” and small skewed results should be expected.

# 6. Conclusion

In this project we successfully implemented a risk-adjusted capitation cost prediction and visualization system. All the steps undertaken from data pre-processing to modeling are adopted from publications related to risk-adjuster systems. This system not only is a tool for predicting next year’s expenditures, but also is an excellent analytical tool for domain experts. Existence of general purpose functions and separated components in the architecture increases scalability and extendability of the system for future growth.

# 

# 

# 

# References

1. Squires, David A. "Explaining high health care spending in the United States: an international comparison of supply, utilization, prices, and quality." *Issue brief (Commonwealth Fund)* 10 (2012): 1-14.
2. Gapenski, Louis C. *Healthcare finance: An introduction to accounting and financial management*. Health Administration Press, 1999.
3. Cumming, Robert B., et al. "A comparative analysis of claims-based methods of health risk assessment for commercial populations." *Final report to the Society of Actuaries* (2002).
4. Ash A, Porell F, Gruenberg L, et al. Adjusting Medicare capitation payments using prior hospitalization data. Health Care Financing Review 1989;10:17–29.
5. Ash A, Porell F, Gruenberg L, et al. An Analysis of Alternative AAPCC Models Using Data from the Continuous Medicare History File. Boston: Brandeis/Boston University, 1986.
6. Ellis RP, Ash A. Refining the diagnostic cost group model: a proposed modification of the AAPCC for HMO reimbursement. Boston: Boston University, 1988.
7. Ellis RP, Ash A. Refinements to the diagnostic cost group (DCG) model. Inquiry 1995;32:418–29.
8. Lamers, Leida M. "Risk-adjusted capitation payments: developing a diagnostic cost groups classification for the Dutch situation." *Health Policy*45.1 (1998): 15-32.
9. Ash, Arlene S., et al. "Using diagnoses to describe populations and predict costs." (2001).
10. Pope, Gregory C., et al. "Risk adjustment of Medicare capitation payments using the CMS-HCC model." (2004).
11. A Colin Cameron and Frank AG Windmeijer. An r- squared measure of goodness of fit for some common nonlinear regression models. *Journal of Econometrics*, 77(2):329–342, 1997.
12. Ash, Arlene S., et al. "Using diagnoses to describe populations and predict costs." (2001).
13. J. R. Quinlan. Learning with continuous classes. In *In Proceedings of AI*, pages 343–348. Adams and Sterling, 1992.
14. Ellis, Randall P. "Risk adjustment in competitive health plan markets."*Handbook of health economics* 1 (2000): 755-845.
15. Arlene S Ash, Randall P Ellis, Gregory C Pope, John Z Ayanian, David W Bates, Helen Burstin, Lisa I Iezzoni, Elizabeth MacKay, and Wei Yu. Using diagnoses to describe populations and predict costs. 2001.
16. Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees*. CRC press, 1984.
17. Jason Brown, Mark Duggan, Ilyana Kuziemko, and William Woolston. Will risk-adjustment decrease health care costs? new evidence from the medicare advantage program. In *Princeton University Working Paper*. 2011.
18. A Colin Cameron and Frank AG Windmeijer. An r- squared measure of goodness of fit for some common nonlinear regression models. *Journal of Econometrics*, 77(2):329–342, 1997.
19. Gregory C Pope, Randall P Ellis, Arlene S Ash, Chuan- Fen Liu, John Z Ayanian, David W Bates, Helen Burstin, Lisa I Iezzoni, and Melvin J Ingber. Principal inpatient diagnostic cost group model for medicare risk adjust- ment. 2001.