Paper 3411-2015

Identifying Factors Associated with High-Cost Patients

Jialuo Cheng, FAIR Health, Inc., New York, NY;

Jeff Dang, Ph.D., FAIR Health, Inc., New York, NY

ABSTRACT

Research has shown that the five percent of patients with the highest individual medical costs account for nearly fifty percent of the total health care expenditures in the United States (NIHCM Foundation, 2012). Using SAS Enterprise Guide and PROC LOGISTIC, a statistical methodology was developed to identify key factors (for example, patient demographics, diagnostic symptoms, co-morbidity, and the type of procedure) associated with the high cost of health care. Analyses were performed using FAIR Health's National Private Insurance Claims (FH NPIC®) database, which contains information on health care utilization and cost in the United States. The analyses focused on chronic conditions, such as Coronary Heart Disease (CHD) and Chronic Obstructive Pulmonary Disease (COPD). Furthermore, heat map and info graphs were created using SAS®i Visual Analytics to illustrate areas with potentially high-cost patients with chronic diseases across the nation.

INTRODUCTION

The accelerating cost of health care poses a growing burden on individuals, businesses and government. Health care spending in the United States has been estimated to be nearly $2.9 trillion in 2013 and accounted for approximately 17% of gross domestic product (GDP) (Centers for Medicare & Medicaid Services, 2014). Interestingly, a majority of health care charges can be attributed to a relatively small number of patients. In fact, research has suggested that about 5% of the U.S. population accounted for nearly half of all U.S. health care spending in 2009 (NIHCM Foundation, 2012).

Research has shown that chronic conditions have contributed greatly to rising health care costs. The Centers for Disease Control and Prevention (CDC) has reported that in 2010 overall spending for cancer totaled $157 billion and that there was a substantial increase in the cost of drugs that are being used to treat cancer (National Cancer Institute, 2011). In 2010, the overall cost for the diagnosis, treatment, and care of heart disease and stroke was estimated to be $193.4 billion (Go, Mozaffarian, Roger, Blaha, & Dai, 2013). Similarly, the spending related to Alzheimer's and other dementias is projected to increase to $189 billion by 2015 (Alzheimer's Association, 2007). Furthermore, the costs associated with diabetes have been estimated to be as high as $176 billion (American Diabetes Association, 2014) and the CDC predicts the amount will increase to $192 billion by 2020 (American Diabetes Association, 2014). Finally, costs are projected to be $49.9 billion (Centers for Disease Control and Prevention, 2012) for kidney and lung related diseases, such as asthma and Chronic Obstructive Pulmonary Disease (COPD).

The primary objective of the research described in this paper is to identify factors associated with high health care charges. A logical step in the management of high-cost conditions is to examine the features that are correlated with high medical charges, such as diagnoses, age and gender. We utilized FAIR Health's National Private Insurance Claims (FH NPIC®) database to learn as much as possible about the characteristics of those patients. The FH NPIC database categorizes services and procedures according to CPT codes[ii]. Then we performed logistic regression via PROC LOGISTIC in SAS Enterprise Guide to examine the relationship between underlying exposures and high expenditures.

METHODS

Data Source

FAIR Health, Inc. houses the nation's largest independent collection of private medical claims data --the FAIR Health National Private Insurance Claims (FH NPIC®) database. The data, which include claims data collected from payors covering more than 150 million privately insured individuals and over 18 billion billed services from 2002 to the present, provide the opportunity for research on a variety of topics, which can deliver actionable results for health care staff and policy makers (FAIR Health, 2014).

Our dataset for this study contains health care claims for services rendered in the United States with dates of service from January 1, 2010 to December 31, 2012. Gender, age, regional variables (North East, West, South and Mid-West regions) and type of health care insurance plan such as Preferred Provider Organization (PPO) and Health Maintenance Organization (HMO) were selected for analysis. In addition, diagnostic conditions were identified by primary and secondary International Classification of Diseases, 9th Revision (ICD-9) codes included in the claim lines. The data used for the study were de-identified.

Identification of Chronic Disease

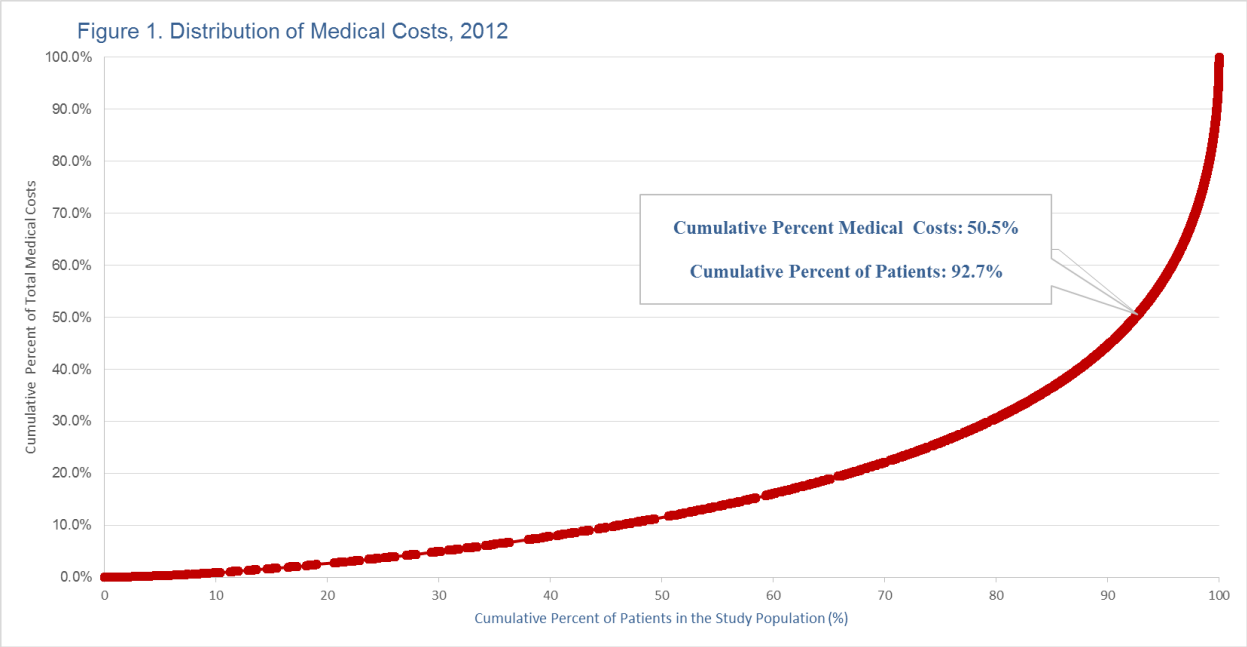
The primary goal was to identify conditions that were significantly correlated with high medical expenditures. By calculating average and total annual charges associated with primary diagnostic conditions, we recognized that there are several chronic conditions, of different prevalence, associated with highest percentile ranges of charges. Focus was on common chronic conditions that are available in the CMS Chronic Condition Warehouse (CCW) research files. We also studied additional chronic conditions with the highest average/total charges in the dataset. This led to considering the 15 primary chronic conditions listed in Table 1. Other medical conditions were listed as "Others" in the Medical Conditions table. If chronic conditions were simultaneously present, then the claim was classified as Comorbidity.

Definitions of High-Cost Patients

Since we targeted persistently high-cost patients based on billing charges, we investigated the distribution of health care charges for each year in our dataset. Our analysis found that the distribution health of care charges is highly concentrated among a small number of patients. Figure 1 shows that less than 10% of

patients consumed more than half of the Medical charges during Year 2012. Therefore, a dichotomous variable was created by stratifying claim lines into one of two categories: the top 10% of the claim lines with the highest charges were designated as high-cost (Farley, Harley, & Devine, 2006) whereas the rest were designated as low-cost claim lines.

Figure 1 Distribution of Medical Charges in Year 2012.



Statistical Analysis

We performed a logistic regression to determine whether selected variables were associated with occurrences of high expenditures as binary dependent variables (high cost or low cost). The independent variables included regional and demographic variables such as age, gender, type of health insurance (or no insurance), and medical condition. We used a logistic regression to evaluate the effects of comorbidity on the medical charges. Finally, we identified procedures related to chronic conditions that are associated with high-cost conditions.

STUDY RESULTS

Characteristics of Medical Charges

Medical charges in the study ranged from $150 to $471,000 with a standard deviation of $1,811. The mean across all claim lines in the dataset was $669 and the median was $312. The difference between mean and median suggested a highly skewed distribution of medical charges.

The SAS Macro to obtain the Descriptive Statistics is given below:

```
%MACRO DESCR_STAT (DATASOURCE, VARIABLE);
        TITLE "MEDICAL COST ANALYSIS BY &VARIABLE ";

        PROC MEANS DATA=&DATASOURCE N MEAN SUM STDDEV MIN MAX
MAXDEC=0;
                WHERE VALIDATION = 'VALID';
                CLASS &VARIABLE;
                VAR TOTAL_CHARGE;
        RUN;

%MEND;

%DESCR_STAT (Patient_Study_Sample, GENDER);
%DESCR_STAT (Patient_Study_Sample, AGE_GROUP);
%DESCR_STAT (Patient_Study_Sample, REGION);
%DESCR_STAT (Patient_Study_Sample, PLAN_TYPE);
%DESCR_STAT (Patient_Study_Sample, CHRONIC_PRIMARY);
```

Table 1 Descriptive Statistics of Medical Charges by Gender, Age Group, Region, Insurance Plan, and Primary Chronic Conditions

Variables	Percentage of Total (%)	Total Charges ($)	Mean Charges ($)	Standard Deviation ($)
Gender				
Female	43	27450923725	662	1628
Male	57	35542572442	670	2029
Age Group				
1-18	13	5557601478	492	1094
19-30	10	6142703033	641	1507
31-40	12	8027904978	704	1591
41-50	16	10821327003	706	1734
51-60	19	13687426678	733	2435
61-70	15	10069587449	712	1960
71-90	15	7697696513	618	1471
Region				
Mid-West	22	12925678389	647	1584
North East	31	19804430406	695	2106
West	13	8695057428	655	1558
South	34	21568329943	666	1756

Insurance Plan				
HMO	3	1534506635	633	1948
IND	7	3914448299	631	1593
MC	1	348313538	556	1295
OTH	11	6673092434	639	1629
POS	43	27739448276	682	1879
PPO	17	10331802574	660	1809
WC	3	348313538	854	2088
No Insurance	15	12103570872	766	2764
Primary Chronic Conditions *				
Diabetes	25	1372891851	444	690
Alzheimer's disease	1	94252772	604	1629
Arthritis	8	173233262	989	2574
Asthma	7	303737318	360	360
Atrial Fibrillation	4	273513608	575	1005
Cancer **	9	1247384335	1093	3878
Chronic kidney disease	4	384641466	725	1401
Chronic liver disease	1	82808231	646	2344
Chronic Obstructive Pulmonary Disease	5	220629664	392	481
Congestive Heart Failure	8	734482880	564	770
Coronary artery disease	12	1089400435	731	1535
Depression	7	602955756	665	919
Osteoporosis	3	267203465	664	2630
Peripheral vascular disease	3	319125296	966	3671
Stroke	4	337524399	763	1530

* The Total Percent excluded Other Medical Conditions
** Cancer contains all Cancer Indicators in ICD-9

Total Charges, mean charges and standard deviation are shown in Table 1. The Total Medical charges in the study population were $62 billion dollars and the mean cost was $669. There were significant differences ($p < .05$) in mean charges in all the variables except for gender ($p = 0.639$) (See Figure 2). The North East Region reported the highest charges as compared to other regions. In particular, the highest rates were found in New York State (See Figure 3). Preferred Provider Organizations (PPOs) had the most coverage in the study group, however, the PPO mean charges were lower than the mean of the charges reported by Workers Compensation Insurance Plans (WC) ($p = 0.002$) and lower than charges for patients with no insurance coverage ($p < 0.0001$) (See Figure 4). Among the medical conditions, Diabetes accounted for 25% of the primary chronic conditions; total charges were the highest in for patients with Diabetes. But the mean charges associated with Diabetes were significantly lower than the mean charges for other conditions except for Asthma. The mean charges related to Cancer were significantly higher than the mean

for other chronic conditions (See Figure 5). The procedures related to high-cost conditions are discussed in the following section.

Figure 2 Average Medical Charges by Gender and Age.

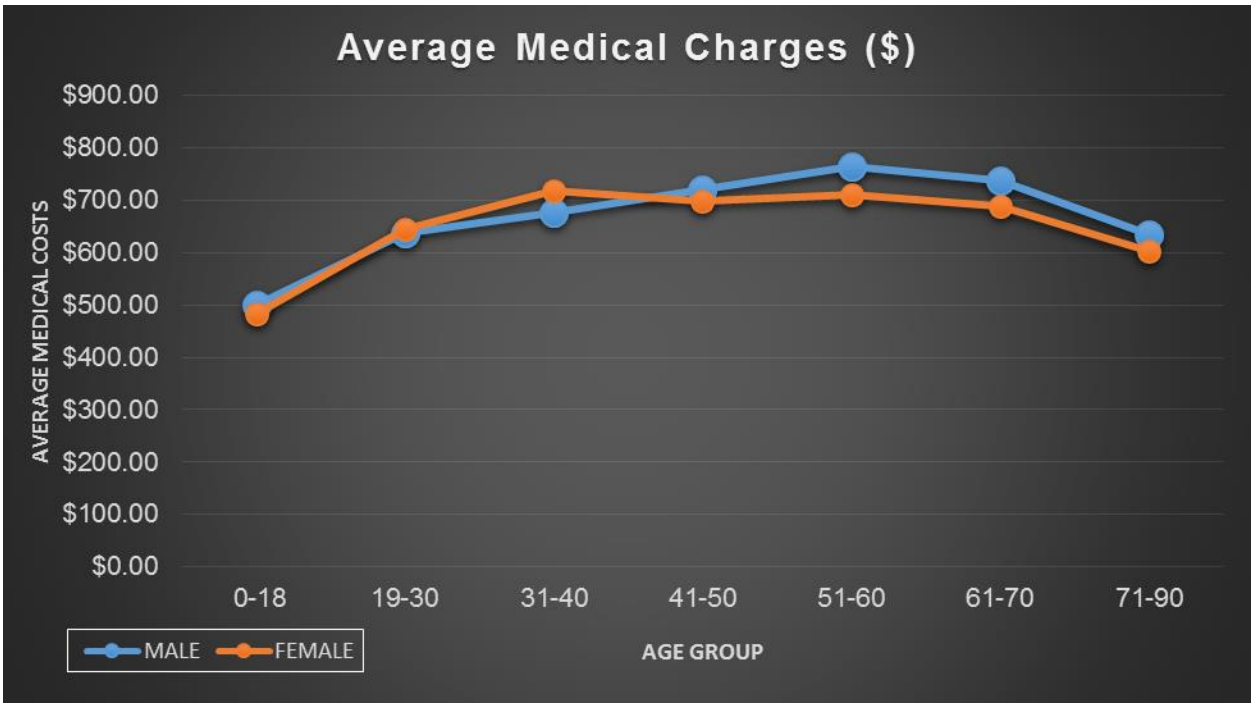


Figure 3 Total Medical Charges by Region.

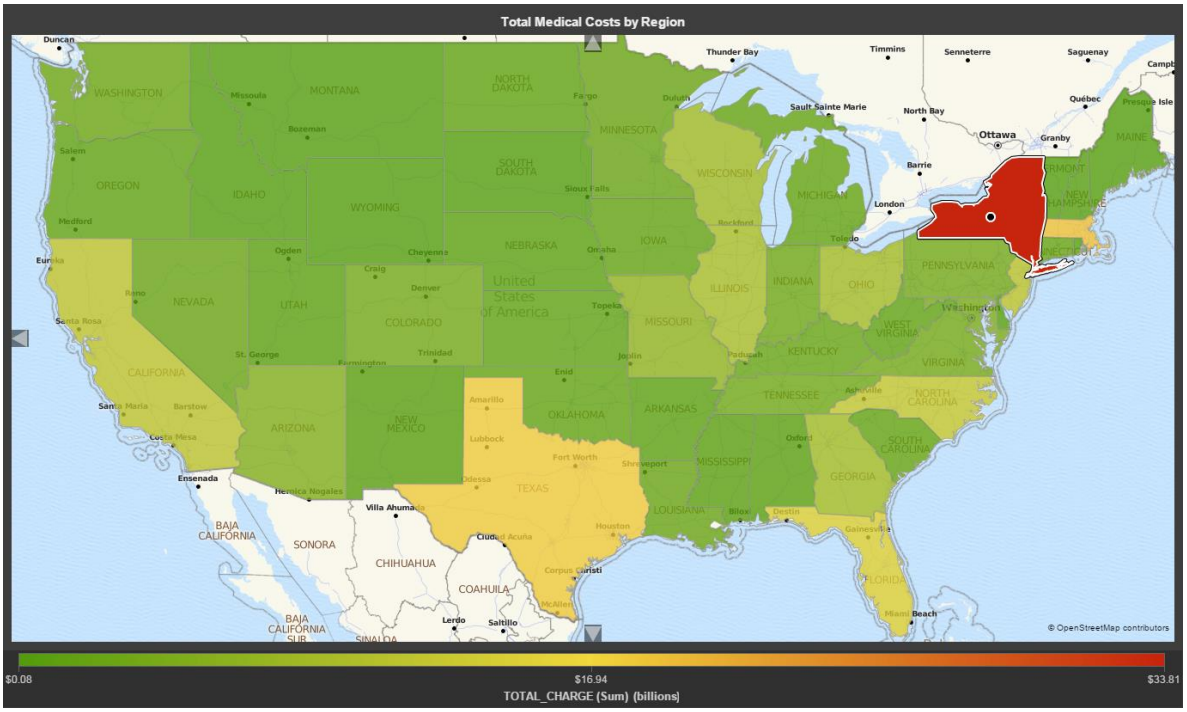


Figure 4 Average Medical Charges by Insurance Plan/No Insurance.

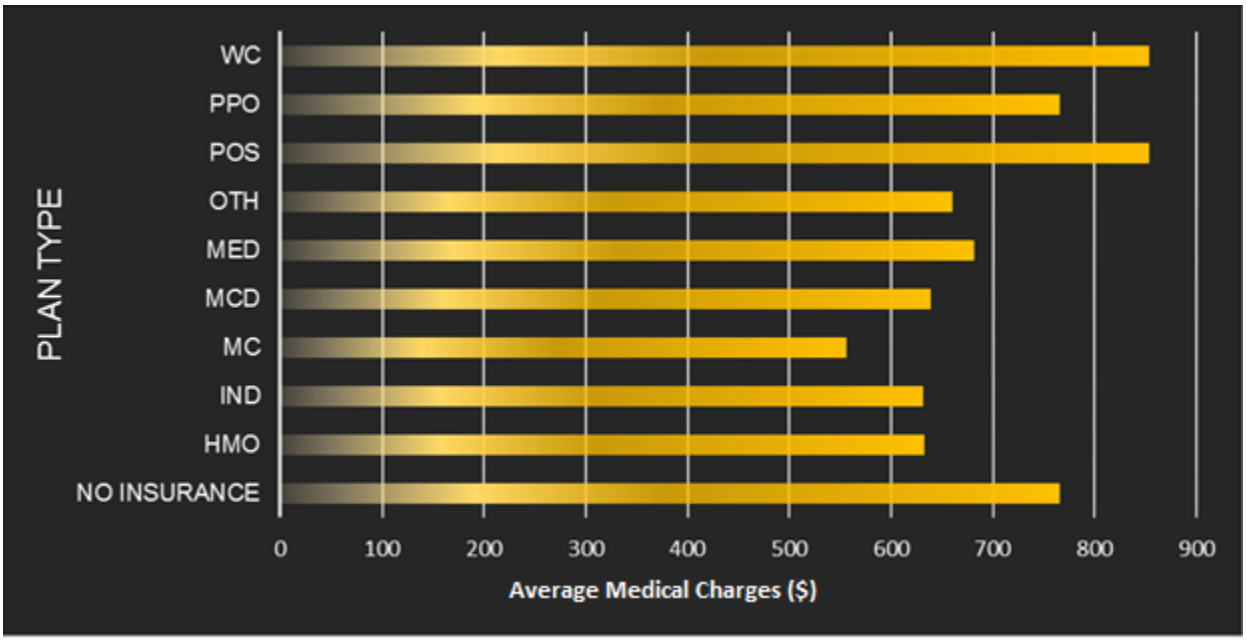
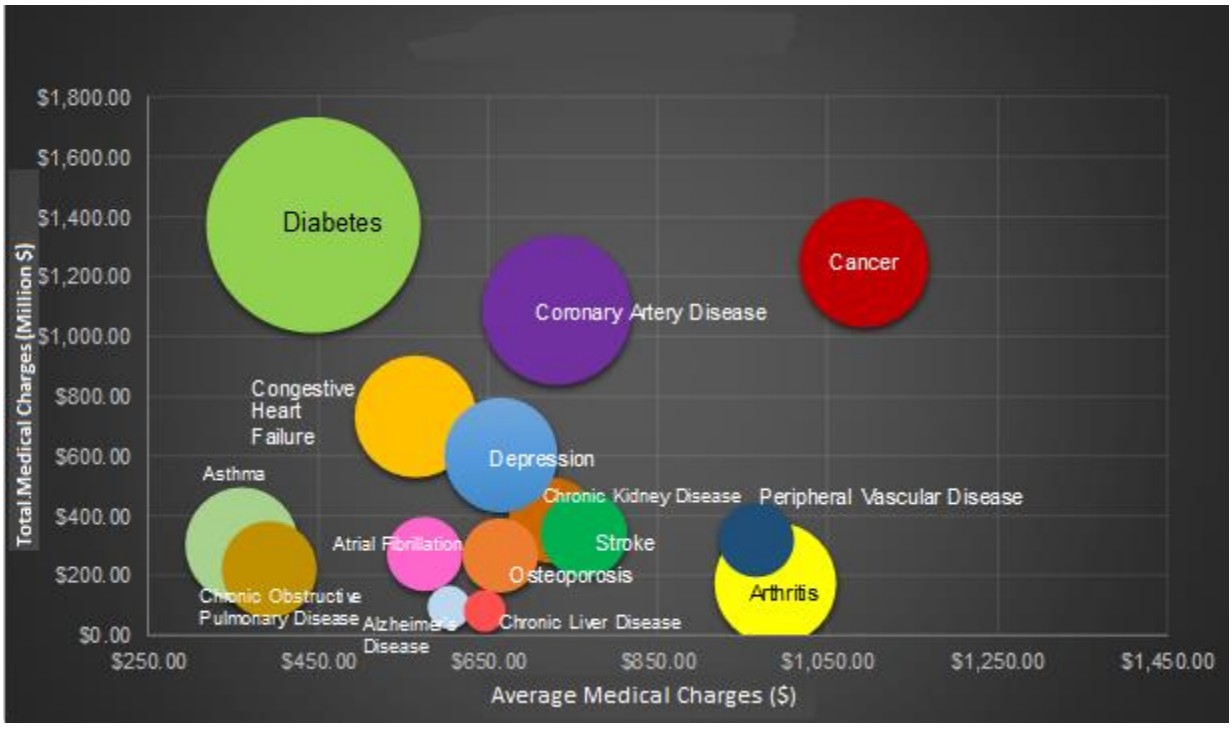


Figure 5 Bubble Charts for Total and Average Medical Charges by Primary Chronic Conditions.



Logistic Regression

According to Table 2 below, age group was statistically significant in the logistic regression model except for the age group from 61 to 70 (p = 0.1324). In general, adults over the age of eighteen were found to have a greater probability of experiencing high medical charges as compared to patients under the age of eighteen. Another significant factor in the logistical regression model was region. It was found that people in the North East region (OR = 1.252, p < .0001), Southern region (OR = 1.112, p = 0.0006) and the West (OR = 1.113, p = 0.0101) were more likely to have high medical bills compared to those in the Mid-West.

With regard to the medical charges of the study group, we used Diabetes as the reference group because it was the most frequently occurring condition in the study areas. The logistic regression analysis indicated that Cancer was the most significant factor affecting medical charges. The odds ratio of Cancer was 5.371, which suggested that the probability of high medical charges was five times higher amongCancer patients as compared to Diabetes patients. Coronary Artery Disease (OR = 4.774, p = 0.037), Arthritis (OR = 4.275, p < .0001), Peripheral Vascular Disease (OR = 4.108, p < .0001), Stroke (OR = 4.028, p < .0001), and Depression (OR = 4.001, p < .0001) were found to have a probability of high medical charges greater than 4 as compared to Diabetes. The odds of having a high medical charge were found to be lower for the following diseases than for Diabetes: Atrial Fibrillation (p = 0.0857), Chronic Liver Disease (p = 0.0680) and Alzheimer's Disease (p = 0.0815). Moreover, the proportion of high-cost claim lines was lower among patients with Asthma (OR = 0.851, p < 0.0001) and Osteoporosis (OR = 0.677, p< 0.0001) than those with Diabetes.

The SAS codes used to perform Logistic regression are shown below:

```
Title "Multiple Logistic Regression analysis of factors associated with high Charges with valid data";

Proc logistic data=Patient_Study_Sample;
        where validation = 'VALID';
        class gender(ref='F')
                region(ref = 'NE')
                insurance
                age_group (ref = '0-18')
                chronic_primary (ref = 'Diabetes');
        model response(event = 'High_Cost') =
                gender
                age_group
                region
                insurance
                chronic_primary;
run;
```

Table 2 Logistic Regression Analysis of factors associated with Gender, Age, Region, Insurance and Primary Chronic Conditions

Independent Variables	Odds Ratio	95% Confidence Interval	p Value
Gender			
Female	**Reference**		
Male	1.061	1.047-1.074	< **.0001**
Age			
0-18	**Reference**		
19-30	1.587	1.540-1.637	< **.0001**
31-40	1.857	1.804-1.912	< **.0001**
41-50	1.871	1.822-1.922	< **.0001**
51-60	1.879	1.828-1.931	< **.0001**
61-70	1.773	1.723-1.824	0.1324
71-90	1.448	1.406-1.491	< **.0001**
Region			
Mid-West	**Reference**		
North East	1.252	1.131-1.274	< **.0001**
South	1.112	1.092-1.132	**0.0006**
West	1.113	1.088-1.136	**0.0101**
Insurance			
Uninsured	**Reference**		
Insured	1.098	1.078-1.118	< **.0001**
Primary Chronic Conditions			
Diabetes	**Reference**		
Alzheimer's disease	2.008	1.559-2.584	0.0815
Arthritis	4.275	3.903-4.674	< **.0001**
Asthma	0.851	0.737-0.981	< **.0001**
Atrial Fibrillation	2.526	2.270-2.812	0.0857
Cancer	5.371	4.931-5.634	< **.0001**
Chronic kidney disease	3.614	3.287-3.974	< **.0001**
Chronic liver disease	2.082	1.703-2.545	0.0680
Chronic Obstructive Pulmonary Disease	1.066	0.919-1.236	< **.0001**
Congestive Heart Failure	2.905	2.552-3.306	< **.0001**
Coronary artery disease	4.774	4.414-5.164	**0.0037**
Depression	4.001	3.692-4.335	< **.0001**
Osteoporosis	0.677	0.497-0.924	< **.0001**
Peripheral vascular disease	4.108	3.652-4.620	< **.0001**
Stroke	4.028	3.591-4.518	< **.0001**

Other Medical Conditions	2.158	2.057-2.264	< **.0001**

Comorbidity Effects

Recent research evidences increasing interest in comorbidity, the co-occurrence to two or more different diseases. (Vogeli, Shields, & Blumenthal, 2007). This is important because the association of two or more than two chronic disorders may increase the likelihood of high medical charges. Table 3 provides the logistic regression results for the comorbidity analysis. The results suggest that a patient with one chronic condition has higher odds of having a high medical charge (OR = 1.087, p < 0.0001) than a patient with no chronic conditions, but the magnitude of the effect was small. Patients with multiple chronic conditions were found to have two times the odds of having high medical charges as compared to those with no chronic conditions.

The Code to execute logistic Regression on Comorbidity Effects is attached:

```
title "Multiple Logistic Regression analysis of factors associated
with high Charges with valid data";

proc logistic data=Comorbidity_Sample;
      where validation = 'VALID';
      class comorbidity(ref = 'No Chronic Condition');
      model response = comorbidity;
run;
```

Table 3 Logistic Regression Analysis of factors associated No Chronic Condition, Only One Chronic Condition and Multiple Chronic Conditions

Independent Variables	Odds Ratio	95% Confidence Interval	p Value
Comorbidity			
No Chronic Condition	**Reference**		
Only One Chronic Condition	1.087	1.002-1.175	< .0001
Multiple Chronic Conditions	2.032	1.884-2.180	< .0001

CONCLUSION

We performed an analysis of claim lines to help identify potential factors associated with high medical charges via PROC LOGISTIC in SAS Enterprise Guide.

Gender and type of insurance were not found to be significant predictors of high medical charges. However, age and region were highly associated with high medical charges. Across all of the chronic diseases that were evaluated in the study, cancer was found to have the highest odds of high medical charges as compared to Diabetes. While patients with Diabetes were less likely to have high medical charges than the

patients with the other major chronic diseases included in the study, the total charges for treating Diabetes were higher than those for other medical conditions because of the high rate (25%) of occurrence of Diabetes in the population.

Moreover, we found that the claims for patients with multiple chronic conditions were more likely to be associated with high charges than claims for patients with no chronic condition or with only one chronic condition. However, the scope of our studies was limited. Without complete classification systems of comorbidity and extensive studies on multiple chronic conditions, it is difficult to establish conclusive cost relationships based on the association of chronic diseases. Despite this limitation, our findings support research literature dealing with the high medical costs associated with comorbidity. (Vogeli, Shields, & Blumenthal, 2007).

References

Agency for Health care Research and Quality. (2014, April 23). *State Turn to Managed Care to Constrain Medicaid Long-Term Care Charges.* Retrieved Februrary 7, 2015, from Agency for Health Care Research and Quality: https://innovations.ahrq.gov/perspectives/states-turn-managed-care-constrain-medicaid-long-term-care-costs

Allison, P. D. (2001). *Logistic Regression Using the SAS System: Theory & Application.* Gary, NC, USA: SAS Institute Inc.

Alzheimer's Association. (2007, May 15). *2007 Alzheimer's Disease Facts and Figures.* Retrieved Februrary 7, 2015, from Alzheimer's Association: http://www.alz.org/wny/documents/fact_sheet_alzheimers_facts_and_figures.pdf

American Diabetes Association. (2014, April 18). *The Cost of Diabetes.* Retrieved Feburary 7, 2015, from American Diabetes Association: http://www.diabetes.org/advocacy/news-events/cost-of-diabetes.html

Billings, J., & Mijanovich, T. (2007, November). Improving The Management Of Care For High-Cost Medicaid Patients. *HealthAffairs, 26*(6), pp. 1643-1655. Retrieved January 11, 2015, from http://content.healthaffairs.org/content/26/6/1643.full.html

Centers for Disease Control and Prevention. (2012, November 23). *Chronic Obstructive Pulmonary Disease among Adults — United States, 2011.* Retrieved Feburary 7, 2015, from Centers for Disease Control and Prevention: http://www.cdc.gov/mmwr/preview/mmwrhtml/mm6146a2.htm

Centers for Medicare & Medicaid Services. (2014, April 2). *National Health Expenditures 2013 Highlights.* Retrieved Feburary 6, 2015, from Centers for Medicare & Medicaid Services: http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/downloads/highlights.pdf

FAIR Health. (2014, July 2). *Announcing the FAIR Health Research Support Program: Featuring the FH NPIC Database.* Retrieved Feburary 7, 2015, from FAIR Health: http://www.fairhealth.org/servlet/servlet.FileDownload?file=01560000000YY0W

Farley, J. F., Harley, C. R., & Devine, J. W. (2006). A Comparison of Comorbidity Measurements to Predict Health care Expenditures. *American Journal of Managed Care 12*, 110-7.

Go, A. S., Mozaffarian, D., Roger, V. L., Blaha, M. J., & Dai, S. (2013, December 18). *Heart Disease and Stroke Statistics—2014 Update.* Retrieved Feburary 5, 2015, from American Heart Association: http://circ.ahajournals.org/content/early/2013/12/18/01.cir.0000441139.02102.80.full.pdf

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2000). *Applied Logistic Regression (2nd Edition).* New York: Wiley.

National Cancer Institute. (2011, January 18). *Cancer Prevalence and Cost of Care Projections*. Retrieved Feburary 7, 2015, from National Cancer Institute: http://costprojections.cancer.gov/

NIHCM Foundation. (2012, July 26). *The Concentration of Health Care Spending*. Retrieved Feburary 5, 2015, from NIHCM Foundation: http://www.nihcm.org/pdf/DataBrief3%20Final.pdf

SAS Institute Inc. (2014). *SAS/STAT 13.2 User's Guide High-Performance Procedures.* Cary, NC: SAS Institute Inc.

Vogeli, C., Shields, A. E., & Blumenthal, D. (2007). Multiple Chronic Conditions: Prevalence, Health Consequences, and Implications for Quality, Care Management, and Costs. *Journal of General Interal Medicine*, 22 (Suppl 3), 391-395.