SOA CASE STUDY

Quantifying Healthcare Industry Adverse Disruptors

Jingjing Xu

Boning Zhang

Jici Zhang

He Zhao

Qian Zhao

Contents

1. Executive Summary

This report discusses the total cost (benefit cost) analysis for a health care corporation within a five year time horizon on a national scale. We provide a quantitative model to identify potential business disruptors and quantify their impacts. Using this model, we provide a prediction of the total cost within five years, conduct a sensitivity analysis on emerging new disruptors, and provide a prediction of capital requirements based on different choices for the probability of defaults.

We earmark six primary influential factors to the total benefit cost: characteristics of diseases (morbidity and average treatment expenditure), average salary in the medical industry, age structure, inflation rate, prevention cost and the size of the insured population. Potential disruptors can be identified as triggers to significant changes in these factors, such as previously uncovered experimental drugs, new technologies, acceleration in population aging, volatility in the inflation rate, etc.

Using appropriate datasets, we implement multivariate multiple regression analysis to obtain relationships between disruptors and benefit cost sources. We then rely on time series analysis for the forecasting process. Based on this model, we can distinguish specific empirical disruptors and quantify their impact.

According to the sensitivity analysis, we evaluate the impact of the six influential factors on the total cost. By this analysis, for any potential disruptor, it is feasible for us to quantify its influence via identifying the relationship between disruptors and influential factors. In other words, impact of emerging disruptor can be quantitatively added to this prediction flexibly.

Additionally, based on national-scale expenditures in the health care industry, our model predicts that the total cost will on average increase at a rate of 1.72% during the next five years. The economic capital requirement for 2016, based on 5%, 1%, and 0.1% PD, should be 101.87%, 102.75%, and 103.75% of the predicted total cost.

2. Purpose and Background

a. Background

The United States as a nation spends tremendous amount on health care, reaching $3.0 trillion in total ($9,523 per person) in 2014. This amount accounted for 17.5% of the nation's Gross Domestic Product and is projected to increase in the future. While this appears to be good news for a health care provider, due to demographic changes, economic growth and other factors, there are still many challenges ahead of stakeholders in the health care industry. As a result, it's necessary to think ahead so as to minimize losses and secure a provider's ongoing operation and development.

In order to maintain the position and to thrive in the health care industry, we are supposed to assess risks that would have a negative influence on our corporation's financial condition. Distinguishing those disruptors and quantifying their impact will help us prepare for possible changes and conceive appropriate strategies in the corporation's operation. For this purpose, we conduct research and generate this comprehensive report on the health care industry's disruptors in the next five years.

b. Related analyses and findings from the literature review

Various health care related aspects such as expenditures, diseases, and risk drivers have been discussed in existing research. Many researchers have been studying them with a variety of methods. According to the North American Industry Classification System (NAICS), the Center for Medicare & Medicaid Service (CMS) breaks down the total National Health Expenditure into different sources[1]. For instance, Goetzel (1998) estimates the impact of ten health risk behavior factors on health care expenditures with two multivariate analysis (logistic and linear regression models) using data of 46,026 employees from a health care purchaser. He controlled for other measured risk and demographic factors to get more precise estimates[2]. Folkerts et al. (1990) implemented

[1] National Health Expenditure Accounts: Methodology Paper, 2014 Definitions, Sources, and Methods

[2] Goetzel, R. Z., Anderson, D. R., Whitmer, R. W., Ozminkowski, R. J., Dunn, R. L., Wasserman, J., & Health Enhancement Research Organization (HERO) Research Committee. (1998). The relationship between modifiable health risks and health care expenditures:

cluster analysis to help clinical chemical diagnosis of liver diseases by applying a hierarchical algorithm for ascertaining a starting partition, followed by the *k*-means algorithm[3].

C. Key Variables

In general, benefit costs can be categorized into four major medical service categories: inpatient services, outpatient services, physician services, and pharmaceuticals. Before determining related disruptors, we explicitly define these four categories as below:

- **Inpatient Services**: benefit costs for patients whose condition requires admission to a hospital, staying overnight or for an indeterminate time;
- **Outpatient Services**: benefit costs for patients who are hospitalized for less than 24 hours;
- **Physician Services**: benefit costs charged by an individual licensed under state law to practice medicine or osteopathy;
- **Pharmaceuticals:** costs related to the supply of drugs, medicines and appliances prescribed by practitioners.

Based on each category's definition, we propose two sources of influence of these costs:

- **N**: size of the insured population;
- **C**: average treatment expenditure.

Moreover, we suggest the following six influential factors according to the sources:

- Characteristic of diseases (incidence/prevalence rate and average treatment expenditure);
- Average salary in medical industry;
- Age structure;
- Inflation rate;
- Prevention cost;
- Insured population.

The correspondence between sources and influential factors are displayed in the following table (Table 1.1):

Source	Influential Factors	Reason
(N)	Characteristic of diseases (Morbidity)	Differences in morbidity will cause the number of patients in the covered population to change.
	Age structure	A change in age structure will result in a change in morbidity. For example, an aging population will lead to an increase in morbidity, which will bring out an increase in the percentage of patients.
	Insured population	Direct indicator.
(C)	Characteristic of Diseases (Average Treatment Expenditure)	Difference in average treatment expenditure for specific diseases will lead to the change in total cost, ceteris paribus.
	Average Salary in Medical Industry	Difference in average salary in medical industry will lead to the change in total cost, ceteris paribus.
	Prevention Cost	Change in prevention cost will lead to change in total cost.
	Inflation Rate	Inflation rate will affect the general level of prices.

Table 1.1 Influential Factors

As these influential factors determine a significant portion of the variation in total cost, volatility in these factors themselves will have a severe impact on the total cost. Here we regard "disruptors" as the unexpected changes in the influential factors. Therefore, they are listed as all the possible reasons that will lead to changes we could not forecast at present. For instance, costs from previously uncovered experimental drugs will result in an unexpected change in the characteristic of related diseases.

---

an analysis of the multi-employer HERO health risk and cost database. Journal of Occupational and Environmental Medicine, 40(10), 843-854.

[3] Folkerts, U., Nagel, D., & Vogt, W. (1990). The use of cluster analysis in clinical chemical diagnosis of liver diseases. Clinical Chemistry and Laboratory Medicine, 28(6), 399-406.

Here we list several ordinary disruptors in the following chart (Table 1.2):

Influential Factors		Potential Disruptors	Impact Duration	
			Long Term	Short Term
Characteristic of Diseases	Morbidity	New Vaccine	√	
		Life Style	√	
		Gene Mutation		√
		Disease Outbreak		√
	Average Treatment Expenditure	New Drugs	√	
		New Facilities	√	
Average Salary in Medical Industry		Market Demand and Supply		√
		Skill and Quality of Physician	√	
Age Structure		Longevity Risk	√	
Insured Population		Business Development	√	
		Government Policy		√
Prevention Cost		New Vaccine	√	
		New Diagnosis Test	√	
Inflation Rate		Monetary Policy		√
		Economy Cycle	√	
		Personal Income		√

Table 1.2 Influential Factors and Potential Disruptors

Since we need to find available data to build a reasonable model for estimation of the disruptor's impact and forecast, we reclassify the total benefit costs as the following six categories based on historical National Health Expenditure Data from Centers for Medicare and Medicaid Services (CMS):

- Hospital care;
- Professional services;
- Other health, residential and personal care;
- Home health care;
- Nursing care facilities and continuing care retirement communities;
- Retail outlet sales of medical products.

Hence, we set the key variables as discussed**. Independent variables** are hospital care; professional services; other health, residential and personal care; home health care; nursing care facilities and continuing care retirement communities; and retail outlet sales of medical products. **Dependent variables** are characteristic of diseases (morbidity and average treatment expenditure); average salary in medical industry; age structure; inflation rate; prevention cost; and insured population.

The relationship between these variables is shown in the following graph (Figure 2.1):

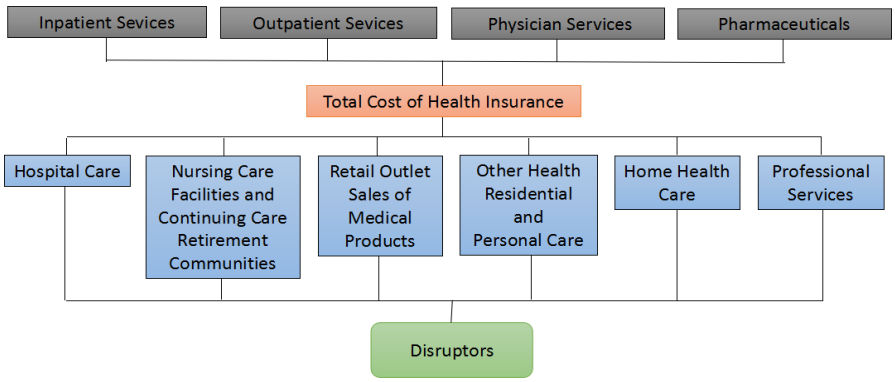

Figure 2.1 Relationship of Variables

For the independent variable characteristic of diseases (morbidity and average treatment expenditure), we adopt clustering analysis to transfer them into categorical variables, which will be further discussed later.

d. Creatively build on existing approaches

Within our report, in order to accurately estimate the total cost on health insurance, we want to determine influential factors that have significant impact on total cost as holistically as possible. However, this is a difficult objective since some important factors are unobservable, and therefore difficult to be quantified. Thus, in our prediction models, we creatively add in some variables that could reflect the impact of those unobservable influence.

Specifically, in order to measure the impact of diseases with different treatment cost and morbidity, we apply cluster analysis in our models, categorizing all diseases into several different clusters that represents the influence of individual disease on total cost. This special implementation allows us to measure the impact from disease on total cost caused by many unexpected events, such as the development of a certain vaccine and a disease burst. We catch the impact in a quantitative way since we assume these type of events would trigger a “cluster jump” effect between diseases. Once this effect happens on a disease, meaning this disease jumps from one cluster to another, the contribution of that disease to total cost could be represented by the difference between two parameters, which quantifies the impact of each cluster on total cost, in our models.

Additionally, we divide total cost into six subclasses by different cost sources: HC (Hospital Care), PS (Professional Services), ORPC (Other Health, Residential, and Personal Care), HHC (Home Health Care), NCF (Nursing Care Facilities and Continuing Care Retirement Communities), and ROS (Retail Outlet Sales of Medical Products). This classification is more convenient for our analysis, since these six subclasses cover all cost resource, and simultaneously, they have almost no overlap between any two of them. Moreover, these six subclasses have corresponding connections with the four major medical services mentioned in this case: inpatient services, outpatient services, physician services, and pharmaceutical. At the same time, these six subclasses are more feasible statistically than the four major services since the data of subclasses is accessible.

Finally, the result of this report gives a comprehensive prediction of future total cost, as well as a quantitative risk management through a calculation of economic capital with different credit rates. We make our analysis not only on the predicted value of cost for the next five years, but also on the volatility of potential cost. This implementation allow us to get a confidence interval of future cost, which then alerts the CEO of this insurance company about the amount of capital his company should hold below certain default probability.

3. Data

a. Database Illustration

As our goal is to examine national data repositories that are suitable, we choose the nation scale data in United States. Considering the consistency of time scale of various databases we used, we choose the sample period from 2002 to 2014.

For the data of the six influential factors, we use the historical National Health Expenditure Data from CMS.GOV.[4] For the average expenditure of 12 selected diseases, we obtain the data from the U.S. Department of Health & Human Services, Agency for Healthcare Research and Quality.[5] For the morbidity data of the 12 selected diseases, we use the database on CDC, from the National Center for Health Statistics /Publications and Information Products/Data Briefs.[6] To take the inflation rate into consideration, we discount all the cost data with the average personal income each year. The data source is from U.S. Department of Commerce/ Bureau of Economic Analysis.[7] For population data, we download data from United Census Bureau.[8] For the data of historical prevention expenditure, we use the database from UNICEF.[9] To measure the salary changing condition of physicians, we use the data from US Bureau of Labor Statistics: Employment Cost Index Historical Listing – Volume IV.[10]

b. Data Processing

Since the data should be relevant, thorough and complete, our data processing procedure follows the steps below:

i. **Verification:**
In the very beginning, we need to ensure the accuracy, consistency and justifiability in our data exploration. To verify those datasets we adopt, we obey the following principles: (1) National official data has the highest priority; (2) All datasets should have the same time scale and sampling scale; (3) If the data source is not official, look them up in as many datasets as we can to ensure the accuracy.

ii. **Standardization:**
Since data of one variable might come from different sources, different datasets have various indicators, magnitudes or presenting formats. We standardize those datasets to ensure consistency.

iii. **Data Cleaning (about the missing data):**
Our target is to give a prediction from 2016 to 2020, we set the sample period from 2002 to 2011 and outcome period from 2011 to 2015. Since we could not find the data of the six influential factors in 2015, we use the predicted data generated by our model as a replacement[11].

iv. **Transformation**
Based on the model we construct, we transfer our dataset into proper format for quantification. For example, we transfer disease characteristics into categorical variables. The methodology of data transformation will be further illustrated in the first part of the model description.

---

[4] Table 2 : National Health Expenditures by Type of Expenditure https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/NationalHealthAccountsHistorical.html

[5] Table: Expenditures by medical Condition. Mean Expenses per Person with Care for Selected Conditions by Type of Service: United States. http://meps.ahrq.gov/mepsweb/data_stats/quick_tables_results.jsp?component=1&subcomponent=0&year=-1&tableSeries=2&tableSubSeries=&searchText=&searchMethod=3&startAt=1&sortBy=

[6] http://www.cdc.gov/nchs/products/databriefs.htm

[7] http://www.bea.gov/iTable/iTable.cfm?ReqID=9&step=1#reqid=9&step=1&isuri=1

[8] http://www.census.gov/population/

[9] http://www.unicef.org/supply/index_38554.html

[10] http://www.bls.gov/data/#employment

[11] This method is proposed by Ralph Kimball.

v. Documentation

Here is a data-variable dictionary for further concise illustration (Table 3.1):

Attribute	Description
HC	Hospital Care
PS	Professional Services
ORPC	Other Health, Residential, and Personal Care
HHC	Home Health Care
NCF	Nursing Care Facilities and Continuing Care Retirement Communities
ROS	Retail Outlet Sales of Medical Products
X9A	Age Structure
INFP	Inflation Rate
PHS	Average Salary in Medical Industry
D1	Disease: Category 1
D2	Disease: Category 2
D3	Disease: Category 3
D4	Disease: Category 4
PO	Insured Population
DRUG	Prevention Cost

Table 3.1 Data-Variable Dictionary

4. Methods, Analysis and Model

a. Overall Introduction

i. Purpose

In this section, we provide a quantitative model to identify potential business disruptors and quantify their impacts. Predictions of total cost in the next five years, similarity analysis of key disruptors, and economic capital requirements based on different probability of defaults can also be obtained.

ii. Assumption

In this model, we use a national scale dataset instead of a corporation scale dataset. Hence, we assume that historical national health expenditure can be regarded as the total benefit cost of a corporation. Correspondingly, the population in the United States could be regarded as the insured population.

As discussed in **part 2.c**, we adopt the multivariate multiple regression to fit the relationship between influential factors and parts of cost. As for influential factors themselves, we apply ARIMA time series model to forecast their values in the next five years.

The following graph (Figure 4.1) shows the general process of our model:

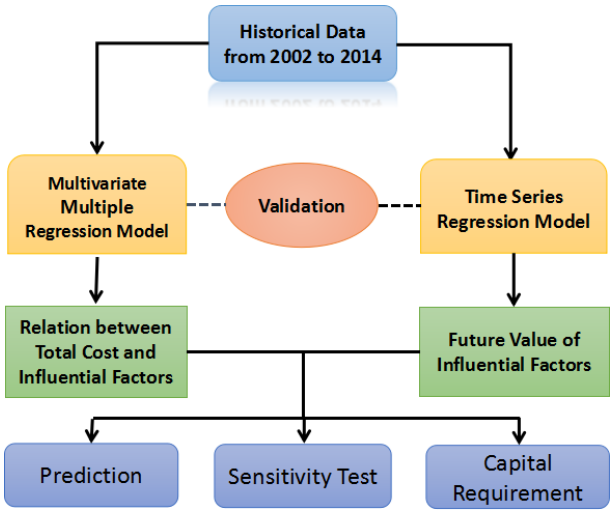

Figure 4.1 General Process

b. Variables Transformation

i. Numerical Data

As the sampling scale might vary in practice, the percentage of changes in six parts of cost is more significant than their original values. Thus, we use logarithms of the five numerical influential factors (which are average salary in medical industry, age structure, inflation rate, prevention cost and insured population) to build the multivariate multiple regression models.

ii. Categorical Data

For the categorical influential factor, characteristic of diseases (morbidity and average treatment expenditure), we adopt a *k*-means clustering method to divide the twelve selected diseases into four groups. Diseases selection is based on their morbidity and average treatment expenditure. We make this selection uniform so that these selected diseases can represent the general feature of disease distribution.

This selected corpus includes: **Cancer; Cataract; Cerebrovascular disease; COPD, asthma; Diabetes mellitus; Gallbladder, pancreatic, and liver disease; Heart conditions; Hyperlipidemia; Hypertension; Intestinal infection; Kidney Disease and Urinary tract infections.**

The two indicators adopted in the *k*-means clustering are (1) morbidity and (2) average treatment expenditure. We can classify all diseases into four clusters using:

$$D_i = \arg\min\left\{(r - r_i)^2 + (c + c_i)^2\right\}$$

The variable, characteristic of diseases, is now quantified as the number of diseases in different diseases clusters. We only adopt three of them in the regression model because independent variables should not be highly correlated. The sum of these categories is fixed so removing one of them is necessary.

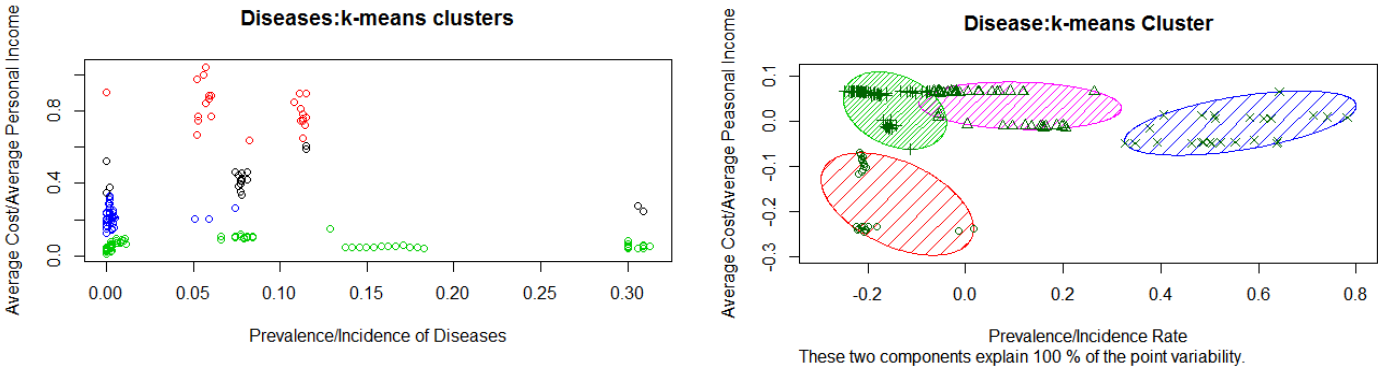

Figure 4.2 Practical coordinates of Diseases within k-means clusters (left)
Scaled results of Diseases within k-means clusters (right)

Figure 4.2 shows the practical coordinates and the scaled result.

The four centers of this *k*-means clustering are listed below (The severity of the impact on the total loss decreases as the Cluster No. grows. (Table 4.1):

Cluster No.	Morbidity	Average Treatment Expenditure
1	0.088961	0.415912
2	0.080793	0.817385
3	0.085223	0.062026
4	0.006911	0.218654

Table 4.1 Four centers of k-means clustering

In conclusion, transformation methodology of all these variables is listed in the following chart: (Table 4.2)

Attribute	Value
lnHC	LN (Hospital Care)
lnPS	LN (Professional Services)
lnORPC	LN (Other Health, Residential, and Personal Care)
lnHHC	LN (Home Health Care)
lnNCF	LN (Nursing Care Facilities and Continuing Care Retirement Communities)
lnROS	LN (Retail Outlet Sales of Medical Products)
lnPHS	LN (Average Salary in Medical Industry)
lnPO	LN (insured population)
lnDRUG	LN (prevention cost)
X9A	90^{th} quantile of people's age
INFP	Inflation rate (percentage)
D1	Number of diseases in Disease Cluster 1
D2	Number of diseases in Disease Cluster 2
D3	Number of diseases in Disease Cluster 3
D4	Number of diseases in Disease Cluster 4 (removed)

Table 4.2 Transformation methodology of variables

c. Methodology

Our model can be divided into two main steps:

- Run a multivariate multiple regression to fit the relationship of six influential factors and six parts of costs;
- Run an ARIMA time series model on the six influential factors to forecast.

i. Multivariate Multiple Regression

This general model is defined as the following equation shows:

$$\begin{cases} \vec{Y} = \beta_{6\times 8}\vec{X} \\ TC = \sum_{i=1}^{6} Y_i \end{cases}$$

The definitions of all letters are listed below: (Table 4.3)

Variable	Definition	Components	Value/Meaning
TC	Total Cost	--	Total Cost
Y	Parts of Total Cost	lnHC	LN (Hospital Care)
		lnPS	LN (Professional Services)
		lnORPC	LN (Other Health, Residential, and Personal Care)
		lnHHC	LN (Home Health Care)
		lnNCF	LN (Nursing Care Facilities and Continuing Care Retirement Communities)
		lnROS	LN (Retail Outlet Sales of Medical Products)
X	Influential factors	X9A	90th quantile of people's age
		INFP	Inflation rate (percentage)
		D1	Number of diseases in Disease Cluster 1
		D2	Number of diseases in Disease Cluster 2
		D3	Number of diseases in Disease Cluster 3
		lnPO	LN (population)
		lnPHS	LN (Average Salary in Medical Industry)
		lnDRUG	LN (prevention cost)
$\boldsymbol{\beta}$	coefficients	β_0	Intercept
		β_{X9A}	If X9A increases 1 unit, Y_i will increase $100\%\beta_{X9A}$ percent
		β_{INFP}	If INFP increases 1 unit, Y_i will increase$100\%\beta_{INFP}$ percent
		β_{lnPHS}	If lnPHS increases 1 percent, Y_i will increase β_{PHS} percent
		β_{D1}	If D1 increases 1 unit, Y_i will increase $100\%\beta_{D1}$ percent
		β_{D2}	If D2 increases 1 unit, Y_i will increase $100\%\beta_{D2}$ percent
		β_{D3}	If D3 increases 1 unit, Y_i will increase $100\%\beta_{D3}$ percent
		β_{lnPO}	If lnPO increases 1 percent, Y_i will increase β_{LNP} percent
		β_{lnDRUG}	If lnDRUG increases 1 percent, Y_i will increase β_{DRUG} percent

Table 4.3 Definitions of all letters

- **1st step: Correlation Analysis**

The correlation coefficients measure the strength of the linear relationship between numerical variables. Since we need to establish a regression model, the examination is necessary: 1) there are no highly correlated independent variables; 2) the correlation coefficient between Y and X is not zero.

The following charts show the correlation coefficients: (Table 4.4)

Correlation Coefficient

	X9A	INFP	lnPHS	D1	D2	D3	lnPO	lnDRUG
X9A	1	-0.2144	0.4177	-0.5276	0.2538	0.5276	0.7268	0.8279
INFP	-0.2144	1	-0.24164	-0.1159	-0.3381	0.3660	-0.2416	-0.2522
lnPHS	0.4177	-0.2416	1	-0.5473	0.5674	0.0488	0.9112	0.8190
D1	-0.5276	-0.1159	-0.54726	1	-0.1852	-0.5568	-0.6162	-0.6350
D2	0.2538	-0.3381	0.567398	-0.1852	1	0.1299	0.5026	0.5126
D3	0.5276	0.3660	0.048842	-0.5568	0.1299	1	0.2523	0.3583
lnPO	0.7268	-0.2416	0.911161	-0.6162	0.5026	0.2524	1	0.9591
lnDRUG	0.8279	-0.2522	0.818978	-0.6350	0.5126	0.3583	0.9592	1

Table 4.4.1 Correlation Coefficient
(Highly correlated variables are filled with dark grey)

Correlation Coefficient

	X9A	INFP	lnPHS	D1	D2	D3	lnPO	lnDRUG
lnHC	0.6775	-0.2408	0.9357	-0.5993	0.5328	0.2215	0.9971	0.9448
lnORPC	0.6697	-0.2593	0.9399	-0.5971	0.5530	0.2135	0.9956	0.9415
lnHHC	0.6561	-0.2371	0.9374	-0.5979	0.4924	0.2078	0.9942	0.9285
lnNCF	0.6874	-0.2498	0.9291	-0.6044	0.5081	0.2090	0.9964	0.9525
lnROS	0.5549	-0.2350	0.9723	-0.5792	0.5648	0.1341	0.9731	0.8990

Table 4.4.2 Correlation Coefficient
(Highly correlated variables are filled with dark grey)

The selection of variables in our following Ordinary Least Squares (OLS) model is based on these coefficients.

- **2nd step: Ordinary Least Squares (OLS)**

The regression model is estimated using ordinary least squares. Ordinary least squares (OLS) is a method for estimating the unknown parameters in a linear regression model, with the goal of minimizing the differences between the observed responses in some arbitrary dataset and the responses predicted by the linear approximation of the data.

According to results of the significant test, indistinctive independent variables are eliminated for each dependent variable. Since D1, D2 and D3 are the results of cluster analysis, all of these three are included in the model if any one is significant. The independent variables included in the six regression models are listed below: (Table 4.5)

Dependent Variables	Independent Variables
HC	X9A, lnPO
PS	X9A, D1, D2, D3, lnPO
ORPC	X9A, D1, D2, D3, lnPO
HHC	X9A, lnINFP, D1, D2, D3, lnPO, lnDRUG
NCF	lnPHS, lnPO
ROS	X9A, D1, D2, D3, lnPO

Table 4.5 Independent variables in the regression models

ii. Time series analysis (Non-seasonal ARIMA models)

To predict the future benefit cost within a five-year time horizon, the predictive value of X is determined with time series analysis. We assume D1, D2, and D3 will not change in five years. Given a time series of data X_t, the autoregressive integrated moving average (ARIMA) model is a tool for understanding and predicting future values in this series. Based on the data we have, the models are established as follows:

X9A	lnPHS	lnPO	lnDRUG	lnINFP
ARIMA(0,2,0)	ARIMA(1,2,0)	ARIMA(0,2,0)	ARIMA(0,1,0)	ARIMA(0,0,0)

Table 4.6 Established Model

d. Accuracy Validation

We have built the multivariate multiple regression model and time series regression models with historical data from 2002 to 2014. We try to use these models to forecast total cost in the future. However, before we use it, it is necessary and important to evaluate the effectiveness of this model, which means we should validate the model before using it.

Since the data we have is from 2002 to 2014, in order to validate our models, we use 10 years data, which from 2002 to 2011, to estimate the parameters of models, and then use the last three years data, which from 2012 to 2014, to test the model's effectiveness, including both the multivariate multiple regressions model and time series models.

For those multivariate multiple regression model (Y_i) which we obtain from the previous section, we predict the values of Y_i, using the independent variable data in 2012, 2013 and 2014. Then, we evaluate the effectiveness of our regression model by employing some indicators. In our report, the indicators and standards that we use are ME (Mean Error), MSE (Mean Square Error), MAD (Mean Absolute Error), and R square, which indicate the scale of regression errors. If these indicators are small enough, it means these regression models are relatively exact and we may use them to make our forecasting; if not, we must modify our models.

Here is the consequence of the effectiveness evaluation for each cost: (Table 4.7)

	ME	MAD	MSE	R-Square
HC	<0.0001	0.004389	<0.0001	0.984432
PS	<0.0001	0.002189	<0.0001	0.994668
ORPC	<0.0001	0.009098	<0.0001	0.936899
HHC	<0.0001	0.002297	<0.0001	0.99422
NCF	<0.0001	0.022013	0.000607	0.708092
ROS	<0.0001	0.004879	<0.0001	0.629138
TC	<0.0001	2.801413	8.861967	0.99822

Table 4.7 Effectiveness evaluation of costs

Figure 4.3 illustrates the estimated values and actual values comparison for each cost.

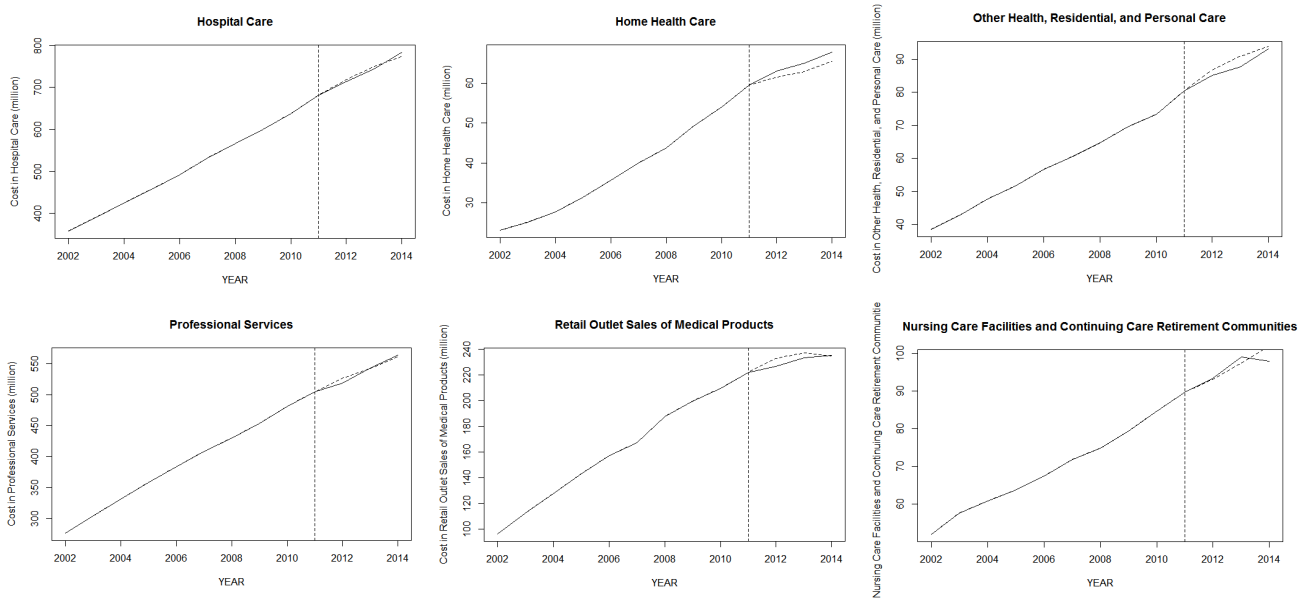

Figure 4.3 Comparison between estimated value and actual value

From the illustrations above, we transparently conclude that the simulated values of each cost in 2012, 2013 and 2014 have the similar trend and scale compared to actual values. In addition, ME, MAD and MSE are small and the R-square for total cost is 0.998, which leads to the conclusion that these multivariate multiple regression models are accepted.

For the time series model for each independent variable, we borrow the estimated regression models from the previous section to carry out our test. Specifically, we use the data of each independent variable from 2002 to 2011, align with the corresponding estimated ARIMA model, to predict the value of 2012, 2013 and 2014. Also, we apply the same method to what we have used in the multivariate multiple regression model to demonstrate the effectiveness of these time series models:

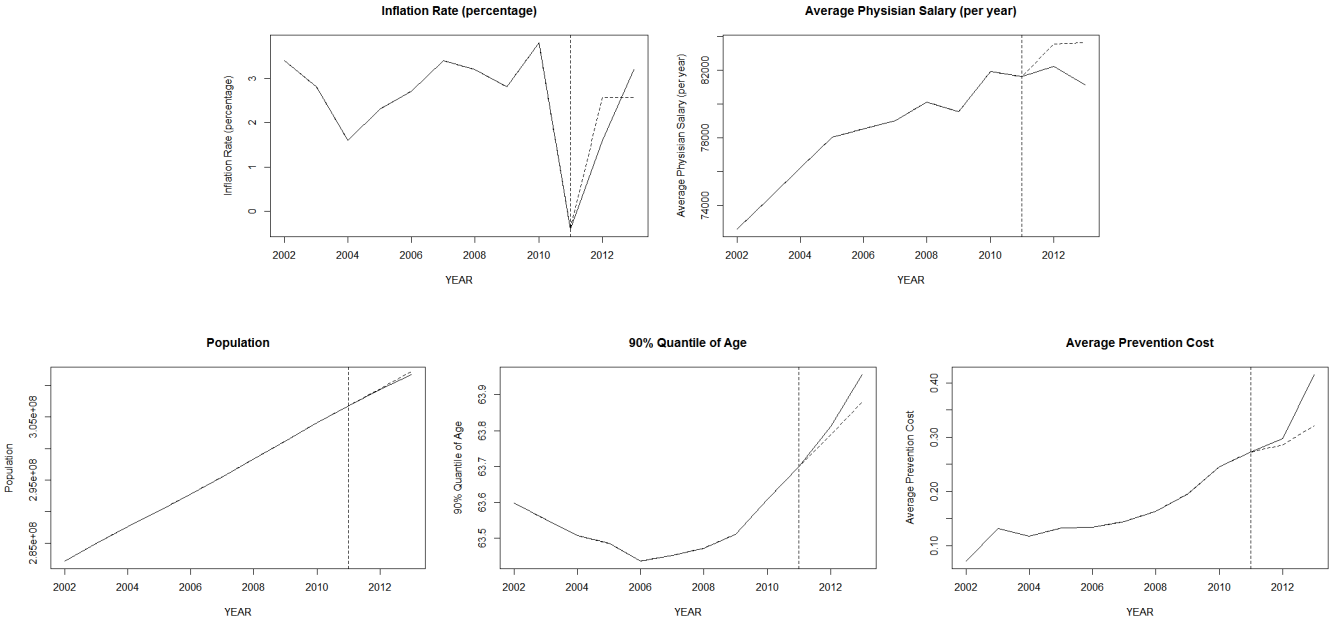

Figure 4.4 Effectiveness of time series models

We notice that the estimated value in the last three years has the same trend with true value, except for physician salary. This salary decline is triggered by some unanticipated reason, and this unexpected salary decline recovered in 2015. Additionally, the small amount of our regression sample also has a negative impact on our estimation of the models' parameters.

According to the analysis, these models are sufficiently accurate to be used within a certain tolerance scope, taking the insufficient historical data into consideration. As a result, we will use them to do our prediction for the next five years in the following sections of this report.

e. Reproducibility

Please find modeling code in the ***Appendix: Codes***.

5. Results, Conclusions and Discussion

a. Results

i. Results of Multivariate Multiple Regression Model

The estimated coefficients of our multivariate multiple regression model are shown below: (Table 5.1)

	lnHC	lnPS	lnORPC	lnHHC	lnNCF	lnROS
Intercept	-137.116	-120.3	-151.5	-236.25	-106.449	-158
X9A	-0.1118	-0.1592	-0.1218	-0.2535	--	-0.3859
INFP	--	--	--	-0.0103	--	--
lnPHS	--	--	--	--	0.6869	--
D1	--	-0.0009	0.0199	0.016	--	-0.0195
D2	--	0.0151	-0.0003	-0.0191	--	0.0234
D3	--	0.0036	-0.0024	0.0323	--	-0.0135
lnPO	7.7148	6.991	8.375	13.11	5.2794	9.623
lnDRUG	--	--	--	-0.0744	--	--
R-square	0.851797	0.862646	0.937909	0.938919	0.853186	0.85322
p-value	<0.05	<0.05	<0.05	<0.05	<0.05	<0.05

Table 5.1 Estimated coefficients of multivariate multiple regression models

It quantifies the total cost based on the six influential factors and predicts possible future benefits cost. For example, under this model, if the data of X9A and lnPO in 2016 could be collected, the cost quantified by lnHC equals to -137.1162+0.1118* X9A+7.7148*lnPO.

b. Conclusions

i. Sensitivity Analysis

In the multivariate multiple regression model, the coefficients β represents the correspondence between influential factors and total benefit cost. For example, if the 90th quantile of age increases 0.1 year, the benefit cost of hospital care will decrease 1.118%.

Furthermore, when quantifying the variable-disease, this report follows two steps:

- Cluster Analysis: assign the disease into one of the four clusters by choosing the minimum distance between cluster centers and disease indicators (morbidity and average treatment expenditure).

- Cluster Jump: if the disease jumps from one cluster to another, there is a "cluster jump" which results in changes in the related dependent variables.

  For example, Disease-A is initially included in Disease Cluster 1, but this disease jumps into Disease Cluster 2 because of the morbidity changes or the average cost change. Then D1 decreases by 1 unit and D2 increases by 1 unit, which results in a 0.016% (0.0151%- (-0.0009%)) increment in the total benefit cost.

Since our aim is to measure the impact of disruptors on total benefit costs, we need to quantify the impact of cost from every category on the total cost. In this report, the weight of each category is calculated using ARIMA time series model, according to the result from regression model (Table 5.2).

Percentage in total benefit costs						
YEAR	**HC**	**PS**	**ORPC**	**HHC**	**NCF**	**ROS**
2016	41.9823%	30.7792%	4.8451%	3.5622%	5.6593%	13.1749%
2017	42.0234%	30.8163%	4.9409%	3.5812%	5.5844%	13.0582%
2018	42.1622%	30.8419%	4.9337%	3.5636%	5.5601%	12.9444%
2019	42.1594%	30.8808%	4.8884%	3.5296%	5.5964%	12.9528%
2020	42.0886%	30.9224%	4.8771%	3.5100%	5.6189%	12.9919%

Table 5.2 Percentages of each disruptor in total benefit costs

For example, 41.9823% of total cost is spent on hospital care in 2016 and the weight is 42.0886% in 2020 theoretically.

To quantify the impact of every influence, we need two steps:

- **1st Step**: Quantify the impact on each category.
- **2nd Step**: Calculate the weighted sum of the values from the 1st step.

The conclusions are listed as follow:

Sensitivity Analysis								
YEAR	**X9A (Δ=0.1)**	**INFP (Δ=1%)**	**PHS (Δ=1%PHS)**	**D1 (Δ=1)**	**D2 (Δ=1)**	**D3 (Δ=1)**	**PO (Δ=1%PO)**	**DRUG (Δ=1%DRUG)**
2016	**-1.6169%**	**-0.0370%**	**0.0389%**	**-0.1310%**	**0.7043%**	**0.0358%**	**7.8300%**	**-0.0027%**
2017	**-1.6151%**	**-0.0370%**	**0.0384%**	**-0.1270%**	**0.7018%**	**0.0379%**	**7.8311%**	**-0.0027%**
2018	**-1.6122%**	**-0.0370%**	**0.0382%**	**-0.1250%**	**0.6998%**	**0.0389%**	**7.8285%**	**-0.0027%**
2019	**-1.6117%**	**-0.0360%**	**0.0384%**	**-0.1270%**	**0.7013%**	**0.0380%**	**7.8254%**	**-0.0026%**
2020	**-1.6124%**	**-0.0360%**	**0.0386%**	**-0.1280%**	**0.7032%**	**0.0370%**	**7.8243%**	**-0.0026%**

Table 5.3 Sensitivity analysis of each factor

The elements in Table 5.3 represent the changing scales these influential factors will cause. For example, if the 90th quantile of people's ages goes up 0.1 year, the total benefit cost decreases 1.6169% in 2016 consistently.

As a result, all disruptors related to these influential factors, such as new vaccines, change in life styles, emerging drugs and technologies, uncovered diseases, etc. The impact of any disruptor can be quantified with the model we provide.

For example, if there is a new vaccine in 2017 which could affect the clusters structure of diseases and prevention cost, then the impact of this disruptor is a sensitivity function with variables (D1, D2, D3, DRUG) in 2017.

In conclusion, this model provides a general way to quantify the impact of a disruptor and furthermore define the key disruptors.

ii. Prediction

From data sources, the values of independent variables in 2002-2014 are collected. Using ARIMA models and multivariate multiple models, we could estimate the values of independent variables in 2015-2020. Since the prediction is for the next five years, we use the simulated dependent values as the actual values in 2015. The estimated values are shown:

Forecast Year	Total Cost
2016	1900.3860
2017	1934.7748
2018	1968.0864
2019	2000.3878
2020	2034.7766

Table 5.4 The expectation of total benefit cost (million)

In addition to the point estimation of total cost in future five years, we could also easily get several confidence intervals of total cost for each year with varying degrees of confidence:

Year	95% CI		99% CI		99.5% CI	
	lower	upper	lower	upper	lower	upper
2016	1855.236	1947.567	1841.441	1962.841	1833.036	1964.56
2017	1862.291	2012.294	1840.421	2037.677	1827.446	2035.713
2018	1866.786	2079.092	1836.766	2116.191	1818.503	2109.143
2019	1868.505	2148.845	1830.059	2199.279	1805.615	2185.879
2020	1875.56	2213.572	1829.038	2274.114	1789.113	2266.675

Table 5.5 Confidence interval of total cost

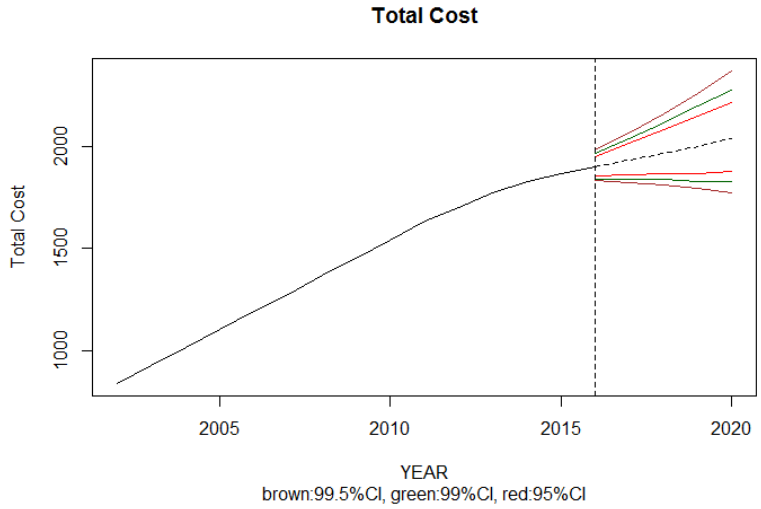

Figure 5.1 Confidence Intervals of total cost

iii. Economic Capital Requirement:

It is necessary for a company to calculate the amount of economic capital required under a certain credit rating to avoid default. We calculate the economic capital requirement by using the quantile of data we collect. In this report, three default probability are chosen: 5%, 1% and 0.1%, since it's meaningful to measure economic capital for a corporation with several different credit rates. As for this specific case, for instance, in 2020, the default probabilities of a corporation will be as much as 5% if the corporation holds at least $2151.0800 million.

Here are the details of economic capital requirement:

Economics Requirement 5% PD	Economics Requirement 1% PD	Economics Requirement 0.1% PD
1935.9767	1952.6341	1971.5854
1989.2421	2016.2352	2047.3456
2042.5053	2081.2065	2125.8465
2096.4835	2148.0672	2208.8443
2151.08	2217.5391	2296.5541

Table 5.6 Economic capital requirement

c. Discussion

Within the previous context, we try to alert the CEO to potential costs in the next five years from a quantitative perspective. We identify the influential degree of each factor on total cost by sensitivity analysis, which allows the health care corporation to scientifically distribute its attention and resources to maintain a sound operation. Keeping track of those disruptors that would lead to an unexpected volatility upon those influential factors is the key objective for the risk management department. The models in this report could be used to estimate how much cost will be incurred due to a certain impulse. Moreover, by the use of time series regression, this report helps the CEO to evaluate the amount of economic capital that the corporation should hold according to different credit ratings. These issues are extremely crucial and useful for a CEO to test the risk exposure and solvency ability of its corporation.

Once we know the exact influential factors and potential disruptors, as well as their sensitivity, some important issues will emerge. From our models, we can eliminate default risk when something unexpected happens by increasing the capital deposit. However, this is not a sustainable method for a company to stably develop its business and build its reputation in the long term. We need to mitigate the potential risks before they happen by some other means, like hedging risk against negative fluctuation on corresponding disruptors and updating the pricing methodology by containing more variables that are leading indicators of potential cost.

Furthermore, although we have solved several meaningful issues in this report, there are still many other topics beyond this report which are critical in the practical field. In this report, we analyze the benefit cost of the corporation, arising from health insurance policy claims. However, in order to control the risk of an insurance company, the CEO should consider potential risk factors from a global perspective. How to evaluate the risk management situation of a corporation is a cutting-edge issue and keeping consistency between risk appetite and risk capability is the core principle to managing risk. There are many other steps that should be taken to enhance the efficiency of a corporation's operation, and then achieve the goals besides predicting the future cost from daily business. Other steps include identifying risk indicators and building risk frameworks.

Finally, all of the conclusions within this report are based on some strict theoretical assumptions, which are too ideal to hold in reality. Hence, the analysis report is meant to give a brief direction on how to quantify the potential risks and cost. We would be extremely pleased if any content in this report could be useful for you. Do not hesitate to communicate with us if you have any suggestions or perspectives about our models and conclusions.

Reference

[1] National Health Expenditure Accounts: Methodology Paper, 2014 Definitions, Sources, and Methods. Centers for Medicare & Medicaid Service.

[2] Goetzel, R. Z., Anderson, D. R., Whitmer, R. W., Ozminkowski, R. J., Dunn, R. L., Wasserman, J., & Health Enhancement Research Organization (HERO) Research Committee. (1998). The relationship between modifiable health risks and health care expenditures: an analysis of the multi-employer HERO health risk and cost database. Journal of Occupational and Environmental Medicine, 40(10), 843-854.

[3]Folkerts, U., Nagel, D., & Vogt, W. (1990). The use of cluster analysis in clinical chemical diagnosis of liver diseases. Clinical Chemistry and Laboratory Medicine, 28(6), 399-406.

[4] National Health Expenditures by Type of Expenditure. https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/NationalHealthAccountsHistorical.html

[5] Expenditures by medical Condition. Mean Expenses per Person with Care for Selected Conditions by Type of Service: United States.
http://meps.ahrq.gov/mepsweb/data_stats/quick_tables_results.jsp?component=1&subcomponent=0&year=1&tableSeries=2&tableSubSeries=&searchText=&searchMethod=3&startAt=1&sortBy=

[6] Average personal income: http://www.cdc.gov/nchs/products/databriefs.htm

[7] Population: http://www.bea.gov/iTable/iTable.cfm?ReqID=9&step=1#reqid=9&step=1&isuri=1

[8] Historical prevention expenditure: http://www.census.gov/population/

[9] US Bureau of Labor Statistics: http://www.unicef.org/supply/index_38554.html

[10] Employment Cost Index Historical Listing: http://www.bls.gov/data/#employment

Appendix: Codes

Software: R

```
#################################################################################################################

#######################          Construction of multivariate multiple regression model
#######################

#################################################################################################################

# This part includes the construction of the multi model between influential factors and disruptors.
#

# 1st Step: Calculate the correlation coefficient of disruptors and influential factors.                       #

# 2st Step: Select significant disruptors for each influential factors based on their correlation              #

# 3st Step: Run a linear regression on these selected factors and revise the model based on the                #

#           regression consequence.                                                                            #

#################################################################################################################


soadata <- read.csv("C:/Users/Jingjing/Desktop/SOA case/soadata.csv")

attach(soadata)

# run a regression model as initial test

group<-t(rbind(X9A,INFP,PI,PHS,D1,D2,D3,LNP,DRUG))

HC_regression<-lm(HC~group)

PS_regression<-lm(PS~group)

ORPC_regression<-lm(ORPC~group)

HHC_regression<-lm(HHC~group)

NCF_regression<-lm(NCF~group)

ROS_regression<-lm(ROS~group)

summary(HC_regression)

summary(PS_regression)

summary(ORPC_regression)

summary(HHC_regression)

summary(NCF_regression)

summary(ROS_regression)


# calculate the correlation coefficient


show(cor(HC,group))

show(cor(PS,group))

show(cor(ORPC,group))

show(cor(HHC,group))
```

```
show(cor(NCF,group))
show(cor(ROS,group))

# construct and revise the model

HCrevise<-lm(HC~X9A+PI+PHS+D1+D2+LNP+DRUG)
summary(HCrevise)
HC_test1<-lm(HC~X9A+D3+INFP+PHS+D1+D2+LNP+DRUG)
PS_test1<-lm(PS~X9A+D3+INFP+PHS+D1+D2+LNP+DRUG)
ORPC_test1<-lm(ORPC~X9A+D3+INFP+PHS+D1+D2+LNP+DRUG)
HHC_test1<-lm(HHC~X9A+D3+INFP+PHS+D1+D2+LNP+DRUG)
NCF_test1<-lm(NCF~X9A+D3+INFP+PHS+D1+D2+LNP+DRUG)
ROS_test1<-lm(ROS~X9A+D3+INFP+PHS+D1+D2+LNP+DRUG)

summary(HC_test1)
summary(PS_test1)
summary(ORPC_test1)
summary(HHC_test1)
summary(NCF_test1)
summary(ROS_test1)

HC_revise<-lm(HC~X9A+LNP)
PS_revise<-lm(PS~X9A+D1+D2+D3+LNP)
ORPC_revise<-lm(ORPC~X9A+D3+D1+D2+LNP)
HHC_revise<-lm(HHC~X9A+D3+INFP+D1+D2+LNP+DRUG)
NCF_revise<-lm(NCF~LNP)
ROS_revise<-lm(ROS~X9A+D3+D1+D2+LNP)

summary(HC_revise)
summary(PS_revise)
summary(ORPC_revise)
summary(HHC_revise)
summary(NCF_revise)
summary(ROS_revise)
```

```
#########################################################################################################################
################################             Time Series Analysis              ################################
#########################################################################################################################
# This part includes the Time Series Analysis of the influential factors and prediction of each factor.        #
#########################################################################################################################
# Time Series Analysis
install.packages("vars")
install.packages("fUnitRoots")
install.packages("forecast")
library("fUnitRoots")
library("forecast")
library("vars"
data_Y<-read.csv("Ydata.csv")
result<- VAR(data_Y,p=1)
data<-read.csv("totaldata.csv")
attach(data)
auto.arima(X9A)
X9A_1<-arima(X9A, order=c(0,2,0))
X9A_2<-forecast(X9A, h=6, level = c(95,99,99.9))
View(X9A_2)
auto.arima(INFP)
INFP_1<-arima(INFP, order=c(0,0,0))
INFP_2<-forecast(INFP, h=6,level = c(95,99,99.9))
View(INFP_2)
auto.arima(PHS)
PHS_1<-arima(PHS, order=c(1,2,0))
PHS_2<-forecast(PHS, h=6,level = c(95,99,99.9))
View(PHS_2)
auto.arima(LNP)
LNP_1<-arima(LNP, order=c(0,2,0))
LNP_2<-forecast(LNP, h=6, level = c(95,99,99.9))
View(LNP_2)
auto.arima(DRUG)
DRUG_1<-arima(DRUG, order=c(0,1,0))
DRUG_2<-forecast(DRUG, h=6, level = c(95,99,99.9))
View(DRUG_2)
```

```
# Prediction of Each Influential Factor
data<-read.csv("trans.csv")
per<-matrix(0,6,6)
initial<-c(0.425391,0.306255,0.050554,0.036816,0.05316,0.127824)
pp<-initial
for (i in 1:6){
  for (j in 1:6){
    per[i,j]=pp%*%data[,j]
  }
  pp<-per[i,]
}
sum(per[6,])

#########################################################################################################
#################################    plot influential factors prediction    ################################
#########################################################################################################

plotHC <- read.csv("C:/Users/Jingjing/Desktop/SOA case/plotHC.csv")
plotPS <- read.csv("C:/Users/Jingjing/Desktop/SOA case/plotPS.csv")
plotORPC<- read.csv("C:/Users/Jingjing/Desktop/SOA case/plotORPC.csv")
plotHHC <- read.csv("C:/Users/Jingjing/Desktop/SOA case/plotHHC.csv")
plotNCF <- read.csv("C:/Users/Jingjing/Desktop/SOA case/plotNCF.csv")
plotROS <- read.csv("C:/Users/Jingjing/Desktop/SOA case/plotROS.csv")

plot(plotHC$YEAR,plotHC$HC,type="l",main="Hospital Care",xlab="YEAR",ylab="Cost in Hospital Care (million)")
lines(plotHC$YEAR,plotHC$preHC,lty=2)
abline(v=2011,lty=2)

plot(plotPS$YEAR,plotPS$PS,type="l",main="Professional Services",xlab="YEAR",ylab="Cost in Professional Services
(million)")
lines(plotPS$YEAR,plotPS$prePS,lty=2)
abline(v=2011,lty=2)

plot(plotORPC$YEAR,plotORPC$ORPC,type="l",main="Other Health, Residential, and Personal
Care",xlab="YEAR",ylab="Cost in Other Health, Residential, and Personal Care (million)")
lines(plotORPC$YEAR,plotORPC$preORPC,lty=2)
abline(v=2011,lty=2)
```

```
plot(plotNCF$YEAR,plotNCF$NCF,type="l",main="Nursing Care Facilities and Continuing Care Retirement
Communities",xlab="YEAR",ylab="Cost in Nursing Care Facilities and Continuing Care Retirement Communities
(million)")

lines(plotNCF$YEAR,plotNCF$preNCF,lty=2)

abline(v=2011,lty=2)


plot(plotHHC$YEAR,plotHHC$HHC,type="l",main="Home Health Care",xlab="YEAR",ylab="Cost in Home Health Care
(million)")

lines(plotHHC$YEAR,plotHHC$preHHC,lty=2)

abline(v=2011,lty=2)


plot(plotROS$YEAR,plotROS$ROS,type="l",main="Retail Outlet Sales of Medical Products",xlab="YEAR",ylab="Cost in
Retail Outlet Sales of Medical Products (million)")

lines(plotROS$YEAR,plotROS$preROS,lty=2)

abline(v=2011,lty=2)


#########################################################################################################

#####################################    plot disruptors prediction      ##################################

#########################################################################################################


plotdisrupters <- read.csv("C:/Users/Jingjing/Desktop/SOA case/plotdisrupters.csv")

disruptergroup<-plotdisrupters[1:12,1:11]

plot(disruptergroup$YEAR,disruptergroup$X9A,type="l",main="90% Quantile of Age",xlab="YEAR",ylab="90% Quantile
of Age")

lines(disruptergroup$YEAR,disruptergroup$preX9A,lty=2)

abline(v=2011,lty=2)

plot(disruptergroup$YEAR,disruptergroup$INFP,type="l",main="Inflation Rate
(percentage)",xlab="YEAR",ylab="Inflation Rate (percentage)")

lines(disruptergroup$YEAR,disruptergroup$preINFP,lty=2)

abline(v=2011,lty=2)

plot(disruptergroup$YEAR,disruptergroup$prePHS,lty=2,type="l",main="Average Physisian Salary (per
year)",xlab="YEAR",ylab="Average Physisian Salary (per year)")

lines(disruptergroup$YEAR,disruptergroup$PHS)

abline(v=2011,lty=2)

plot(disruptergroup$YEAR,disruptergroup$LNP,type="l",main="Population",xlab="YEAR",ylab="Population")

lines(disruptergroup$YEAR,disruptergroup$preLNP,lty=2)

abline(v=2011,lty=2)
```

```
plot(disruptergroup$YEAR,disruptergroup$DRUG,type="l",main="Average Prevention Cost",xlab="YEAR",ylab="Average
Prevention Cost")

lines(disruptergroup$YEAR,disruptergroup$preDRUG,lty=2)

abline(v=2011,lty=2)


################################################################################################################

################################################ CI plots ########################################################

################################################################################################################

#                        Plotting the Confidence Interval of Total Cost prediction                              #

################################################################################################################


CIplot <- read.csv("C:/Users/Jingjing/Desktop/SOA case/soacase/CIplot.csv")

plot(CIplot$YEAR,CIplot$TC999h,type="l",col="brown",main="Total Cost",xlab="YEAR",ylab="Total
Cost",sub="brown:99.9%CI ")

lines(CIplot$YEAR,CIplot$TC999l,col="brown")

lines(CIplot$YEAR,CIplot$TC99h,col="dark green")

lines(CIplot$YEAR,CIplot$TC99l,col="dark green")

lines(CIplot$YEAR,CIplot$TC95h,col="red")

lines(CIplot$YEAR,CIplot$TC95l,col="red")

lines(CIplot$YEAR,CIplot$TC,type="p")

lines(CIplot$YEAR,CIplot$Tcpre,lty=2)

abline(v=2016,lty=2)




################################################################################################################

#                                              Disease Clustering                                              #

################################################################################################################


#k is the clustering number

#yrs: sample time scale (in years)

#dis_n : number of diseases

clustertest3 <- read.csv("C:/Users/Jingjing/Desktop/SOA case/soacase/clustertest3.csv")

k=4

yrs=13

dis_n=12
```

```
cl3<-kmeans(clustertest3,k)
plot(clustertest3, col = cl3$cluster)


clmatrix<-matrix(0,nrow=(yrs),ncol=(k))


for(j in 1:(yrs))
{
for(i in 1:(dis_n))
{
  if((cl3$cluster[(j-1)*12+i])==1)
  {
    clmatrix[j,1]=clmatrix[j,1]+1
  }
  else if(cl3$cluster[(j-1)*12+i]==2)
  {
    clmatrix[j,2]=clmatrix[j,2]+1
  }
  else if(cl3$cluster[(j-1)*12+i]==3)
  {
    clmatrix[j,3]=clmatrix[j,3]+1
  }
  else if(cl3$cluster[(j-1)*12+i]==4)
  {
    clmatrix[j,4]=clmatrix[j,4]+1
  }
  else if(cl3$cluster[(j-1)*12+i]==5)
  {
    clmatrix[j,5]=clmatrix[j,5]+1
  }
  else if(cl3$cluster[(j-1)*12+i]==6)
  {
    clmatrix[j,6]=clmatrix[j,6]+1
  }
}
}
View(cl3$centers)
```