Cost Changes of Diabetes Treatment by Age Group and Region in South Korea

Hyerim Kim
(hrk006@ucsd.edu)
March 16, 2016
1. Introduction

Diabetes has become a significant public health problem in recent years, especially with the growth of the aging population ranging more than 50 years old. The health care cost for diabetes treatment has also increased rapidly, which imposes a considerable burden on patient and national health insurance fund. This research project expects to describe the trends in healthcare cost for diabetes treatment in Korean Healthcare System. This research analyzes public use files provided by the National Health Insurance Sharing Service of South Korea\(^1\). The healthcare insurance system of South Korea is very different from that of the U.S. South Korea has a compulsory healthcare system called “National Health Insurance (NHI) system” and all residents in South Korea are required to enroll the NHI system by law. The National Health Insurance Corporation (NHIC) is the only public insurance institution operated by the Ministry of Health and Welfare in South Korea\(^2\).

Figure 1. Structure of National Health Insurance System in South Korea

Source: NHI Program

---

\(^1\) NHISS (National Health Insurance Sharing Service), http://nhiss.nhis.or.kr/op/up/opup300.do?data_pttn_cd=01

\(^2\) NHI Program, http://www.nhic.or.kr/static/html/wbd/g/a/wbdga0405.html
The target audience of health care research is usually divided into two sectors: business sector and public sector. The value of research for business sector is to provide actionable solutions by predicting customer behavior and developing competitive edge. The value for public sector is to provide sustainable solution by monitoring the safety of the healthcare system and determining appropriate treatment paths. This research project sets public sector managing health care system as target audience and expects to provide a clue for estimating the cost changes for diabetes treatment.

Figure 2. Dataset attributes comparison between business and healthcare sector

<table>
<thead>
<tr>
<th>BUSINESS</th>
<th>VALUE</th>
<th>HEALTHCARE</th>
</tr>
</thead>
<tbody>
<tr>
<td>VOLUME</td>
<td>Providing actionable solutions (e.g., predicting customer behavior developing competitive edge)</td>
<td>Providing sustainable solutions (e.g., monitoring the safety of the healthcare system &amp; determining appropriate treatment paths)</td>
</tr>
<tr>
<td>VARIETY</td>
<td>* The amount of data generated by organizations, individuals or machines</td>
<td>* Data in all forms: traditional, unstructured, semi-structured</td>
</tr>
<tr>
<td>VELOCITY</td>
<td>* Data in all forms: structured, unstructured, semi-structured</td>
<td>* The speed of data generation, delivery or processing</td>
</tr>
<tr>
<td></td>
<td>Providing better understanding of problems (e.g., modeling our climate)</td>
<td>CHALLENGES</td>
</tr>
<tr>
<td></td>
<td>Providing sustainable solutions (e.g., modeling the safety of the healthcare system &amp; determining appropriate treatment paths)</td>
<td>CHALLENGES</td>
</tr>
<tr>
<td></td>
<td>* Data scientists: analysts and statisticians</td>
<td>* Data mining: storing, linking, processing</td>
</tr>
</tbody>
</table>

Source: Jee and Kim (2013)

2. Literature Review

Diabetes is the much-discussed issue in healthcare sector and many experts conduct research in various different perspective. One of the most remarkable research articles is about
global estimates of diabetes prevalence. Guariguata argues that diabetes is a serious and increasing global health burden and estimates of prevalence are essential for appropriate allocation of resources and monitoring of trends. By applying Analytic Hierarchy Process for 219 countries and territories, the research found that most people with diabetes live in low- and middle-income countries and these will experience the greatest increase in cases of diabetes over the next 22 years (Guariguata et al. 2014). Another research also shows that the prevalence of diabetes for all age groups worldwide was estimated to be 2.8% in 2000 and 4.4% in 2030. The prevalence of diabetes is higher in men than women, but there are more women with diabetes than men (Wild and Roglic, 2004). Previous studies have focused on estimation of diabetes prevalence, detecting factors which cause diabetes (Jee et al, 2010),

3. Data

3.1 Data Source

This research uses the National Health Insurance (NHI) claims database, which is derived from a large population-based database. The database contains multiple claims of 1 million patients who visited medical facilities more than once in 2003, 2008, and 2013 respectively. The 1 million patients were selected by conducting random sampling. The database is provided as .csv file (Comma Separated Value), which is clean and comparatively disaggregated enough to conduct data analysis. So far, US and Taiwan also provide similar database to the public: US Nationwide Inpatient Sample (NIS) and Taiwan National Health Insurance Research (NHIRD).
3.2 Variables

The dataset is composed of 17 variables including individual patients’ sex, age, location, type of healthcare facilities (outpatient/inpatient), type of medical treatment, number of days in hospitalization/visit, number of days in medication, and payment information. The dependent variable is total payment (EDEC_TRAMT) and other variables would be used for explanatory variables. Please see the appendix for full description of the variables.

3.3 Constructed Data and Problems

Although this database is comparatively clean, it still needs to be constructed to create new variables. The newly constructed variables are as follows:

MAIN_SICK13: categorical variable only include E100-E140 (codes for diabetes mellitus codes) in MAIN_SICK

AGE13_R2, SEX13_R2, SIDO13_R2, FORM_CD13_R2, DSBJT_CD13_R2: recoded as categorical variables and changed value. For example, 1=male and 2=female.

AGE_R, Facility, Sex, Treatment, Location, TotalPayment, Hospitalization, Medication, Age: mean value of claims of a same patient.

The target database which is directly used for this research is a subset of original database provided by National Health Insurance Sharing Service (NHISS). The target database is constructed by only extracting claims of which sickness is diagnosed as diabetes and contains approximately 287000 observations in 2013.
3.4 Statistical Descriptions

According to the diabetes fact sheet in Korea 2015 provided by Korean Diabetes Association, about 2.7 million Korean people (8.03%) aged 30 years or older had type 2 diabetes in 2013. 8.8% of men and 7.3% of women were diagnosed as diabetes. The prevalence of diabetes (>= 30 years old) increased from 5.6% in 2006 to 8.0% in 2013. 60% of subjects with type 2 diabetes were treated with more than two classes of antidiabetic medication, but drug adherence rate was only 45% in 2013. Only 16.4% of patients with type 2 diabetes were being treated with insulin in 2013. As figure 1 shows, number of patients diagnosed as diabetes significantly increases among people who are older than 40.

Figure 3. Prevalence of Diabetes According to Age and Gender (2013)

---

3 Source of Data: The estimated percentages and the total number of people over the age of 30 with type 2 diabetes were determined using the information from the National Health Information Database from January 2002 through to December 2013 made by National Health Insurance Service (NHIS). Diagnosis of diabetes was based on the disease-classification codes from the health insurance claim forms and database on Health Screening Service. When the database on Health Screening Service was used, diabetes was diagnosed based on fasting glucose (<= 126mg/dL). When the database on health insurance claim forms was used, type 2 diabetes was defined based on ICD-10 code (E11-E14) and prescription of antidiabetic medications.
4. Methods and Quantitative Results

4.1 Data Analysis Methods

This research project mainly focuses on describe cost changes in diabetes treatment. Most of the analysis zooms in exploratory data analysis, and random forest modeling is adopted to estimate cost changes in diabetes treatment. Word equation for the predictive model is as follows:

Type of random forest: regression

\[ Total\ Payment\ per\ Patient = \beta_0 + \beta_1 \times \text{Age} + \beta_2 \times \text{Sex} + \beta_3 \times \text{Medical\ Facility} + \beta_4 \times \text{Type\ of\ Treatment} + \beta_5 \times \text{Location} + \beta_6 \times \text{Number\ of\ days\ in\ Outpatient\ and\ Inpatient} + \beta_7 \times \text{Number\ of\ Days\ in\ Medication} + \epsilon \]

4.2 Exploratory Data Analysis

Figure 4. Costs of Diabetes Treatment by Age Group According to Type of Medical Facilities (won) (2013)
Figure 4 shows two graphs. The first graph describes annual total costs of diabetes treatment by age group and the second graph shows total payment per treated patient by age group. Both graphs contain medical facility information. There are 8 types of medical facilities patients can visit in Korea: hospital, maternity clinic, public health center, psychiatric clinic and psychiatric hospital, and each medical facilities are divided into two outpatient or inpatient.

Generally speaking, hospital is much bigger than public health center and provides various types of treatment. The patients who were received diabetes treatment in 2013 only visited hospital and public health center.

According to the left-side graph, annual total costs for diabetes treatment significantly increased for patients who are older than 50. The age group which has the highest total cost is 70-74 years old in 2013. It is notable that costs for hospitalization comprises the great majority of total costs of patients who are older than 80. By analyzing total payment per treated patient, we can find that total payment for hospitalization is much higher than outpatient treatment. Total payment for using public health center (outpatient) gradually increases from younger age to older age, while total payment for hospital (outpatient) remain almost same or slightly decreases. Total payment for hospital (hospitalization) varies in age distribution, but significantly increases in age group ranging 85 years and older.
As figure 5 shows, payment for diabetes treatment varies according to age group and type of medical treatment. Total payment of patients who are older than 85 is much higher than that of other patients. Total payment per treated patient is high among patients whose age ranges are from 05-20 and from 75 and older. NHIS generally pays 70% of total payment per claim on average and individual patient pays the remainder. The ratio of NHIS payment over total payment is smaller in younger age group, which means younger patients pay more than older patient proportionally. According to the right-side graph, total payment for treatment of plastic surgery is the most expensive, while payment for dermatology is the cheapest.
Figure 6 shows costs changes of diabetes treatment per patient by age group in 2003, 2008 and 2013. There would be statistical noise in age group 1, 2, and 3 because of very small number of patients. However, it is notable that there happened big cost changes between 2003 to 2008 in age group ranging 75 and older.
Regional variation in total costs for diabetes treatment per patient has changed significantly since 2003. The most expensive regions seem to change from Chungnam Province and Gangwon Province in 2003 to Jeonnam Province in 2013. The highest cost for diabetes treatment were found in Jeonnam, Gyeongnam, Chungbuk Province and the lowest cost was found in Jeju island in 2013. Salty recipes in Jeonnam Province might have impacts on diabetes diagnosis.

4.3 Predictive Analytics Model: Random Forest

For predictive analysis, this research project estimated random forest model. Target variable is total payment and exploratory variables are age group, sex, medical facility, type of treatment, location, number of days in outpatient and inpatient, and number of days in medication. It is largely known that random forest algorithm produces more accurate models and is much more robust to changes in the data, that is, it is very robust to statistic noise (Williams, 2011).

Figure 8. Variable Importance of Random Forest Model

By analyzing the database using random forest modeling, we can find that type of medical facilities a patient visited (Facility) is the variable which most impacts on cost changes in diabetes treatment. Whether a patient is outpatient or inpatient would be significantly
associated with cost changes. Number of days in inpatient and outpatient (Hospitalization) is also closely related with cost changes. Age, type of treatment, and age are also important variables.

4.4 Testing the Models

Figure 9. Error Rates

![Error Rates Graph](image)

Approximately 63.21% of variables are explained by the random forest model. As figure 10 shows, the error rate of the overall model generally decreases as each new tree is added to the ensemble. However, the error rate at the beginning part is significantly high.

Figure 10. Predicted vs Observed Plot

![Predicted vs Observed Plot](image)
The Predicted Versus Observed plot is relevant for regression models by predicting a continuous value rather than a discrete value. Figure 10 displays the predicted values against the observed values. Two lines are not fitted so well, which means the predicted values are not the same as the actual observations.

5. Conclusion and Further Research

It is found that this random forest model is not perfectly accurate and has some errors. However, by combining the exploratory data analysis and the random forest model, it is predictable that some patient would cost fairly much of money under certain conditions. For example, if a diabetes patient is older than 70, living in the Southern provinces of South Korea, hospitalized in a hospital which provides comprehensive treatment, received medical treatment such as plastic surgery and orthopedics, we can predict that treatment cost for the patient would be much higher than average cost.

Lack of detail data on patient information such as height, weight, blood pressure, smoking status, limits the scope of research methodology. Lack of time is also an important factor to hinder conducting more complete and productive research. The scope for further research topic is comparatively broad. Given that this database is the cohort data including 12-year-long records, researchers could conduct various trend analysis. For example, health care cost inflation, trends in cancer incidence, geographic variation in treatment and so on. It is also worth consideration to conduct comparative analysis between US, South Korea and Taiwan.
Bibliography


# Appendix

Samples of the raw data

---

**General information table**

**Diagnostic code table**

**Inpatient medical history table**

**Prescription number**

**Outpatient prescription table**

---

Souce: Kim and Jin, 2013
### Full description of the variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>STND_Y</td>
<td>Year</td>
</tr>
<tr>
<td>IDV_ID</td>
<td>Patient’s ID</td>
</tr>
<tr>
<td>SEX</td>
<td>1 = male, 2 = female</td>
</tr>
<tr>
<td>AGE_GROUP</td>
<td>Group 1 = 0~4 years old</td>
</tr>
<tr>
<td></td>
<td>Group 2 = 5~9 years old</td>
</tr>
<tr>
<td></td>
<td>Group 3 = 10~14 years old</td>
</tr>
<tr>
<td></td>
<td>Group 4 = 15~19 years old</td>
</tr>
<tr>
<td></td>
<td>Group 5 = 20~24 years old</td>
</tr>
<tr>
<td></td>
<td>Group 6 = 25~29 years old</td>
</tr>
<tr>
<td></td>
<td>Group 7 = 30~34 years old</td>
</tr>
<tr>
<td></td>
<td>Group 8 = 35~39 years old</td>
</tr>
<tr>
<td></td>
<td>Group 9 = 40~44 years old</td>
</tr>
<tr>
<td></td>
<td>Group 10 = 45~49 years old</td>
</tr>
<tr>
<td></td>
<td>Group 11 = 50~54 years old</td>
</tr>
<tr>
<td></td>
<td>Group 12 = 55~59 years old</td>
</tr>
<tr>
<td></td>
<td>Group 13 = 60~64 years old</td>
</tr>
<tr>
<td></td>
<td>Group 14 = 65~69 years old</td>
</tr>
<tr>
<td></td>
<td>Group 15 = 70~74 years old</td>
</tr>
<tr>
<td></td>
<td>Group 16 = 75~79 years old</td>
</tr>
<tr>
<td></td>
<td>Group 17 = 80~84 years old</td>
</tr>
<tr>
<td></td>
<td>Group 18 = 85~ years old</td>
</tr>
<tr>
<td></td>
<td>Group 19 = 86~ years old</td>
</tr>
<tr>
<td>SIDO</td>
<td>Patient’s location</td>
</tr>
<tr>
<td></td>
<td>11 = Seoul City</td>
</tr>
<tr>
<td></td>
<td>26 = Busan City</td>
</tr>
<tr>
<td></td>
<td>27 = Daegu City</td>
</tr>
<tr>
<td></td>
<td>28 = Incheon City</td>
</tr>
<tr>
<td></td>
<td>29 = Gwangju City</td>
</tr>
<tr>
<td></td>
<td>30 = Daejeon City</td>
</tr>
<tr>
<td></td>
<td>31 = Ulsan City</td>
</tr>
<tr>
<td></td>
<td>36 = Sejong City</td>
</tr>
<tr>
<td></td>
<td>41 = Gyeonggi Province</td>
</tr>
<tr>
<td></td>
<td>42 = Gangwon Province</td>
</tr>
<tr>
<td></td>
<td>43 = Chungcheongbuk Province</td>
</tr>
<tr>
<td></td>
<td>44 = Chungcheongnam Province</td>
</tr>
<tr>
<td></td>
<td>45 = Jeollabub Province</td>
</tr>
<tr>
<td></td>
<td>46 = Jeollanam Province</td>
</tr>
<tr>
<td></td>
<td>47 = Gyeongsangbuk Province</td>
</tr>
<tr>
<td></td>
<td>48 = Gyeongsangnam Province</td>
</tr>
<tr>
<td></td>
<td>49 = Jeju Province</td>
</tr>
<tr>
<td>RECU_FR_DT</td>
<td>Starting date of visiting medial facilities</td>
</tr>
<tr>
<td>FORM_CD</td>
<td>Type of facilities the patient visited</td>
</tr>
<tr>
<td></td>
<td>02 = hospital (hospitalization)</td>
</tr>
<tr>
<td></td>
<td>03 = hospital (outpatient)</td>
</tr>
<tr>
<td></td>
<td>06 = maternity clinic (hospitalization)</td>
</tr>
<tr>
<td></td>
<td>07 = public health center (hospitalization)</td>
</tr>
<tr>
<td></td>
<td>08 = public health center (outpatient)</td>
</tr>
<tr>
<td></td>
<td>09 = psychiatric clinic (daytime ward)</td>
</tr>
<tr>
<td></td>
<td>10 = psychiatric clinic (hospitalization)</td>
</tr>
<tr>
<td></td>
<td>11 = psychiatric hospital (outpatient)</td>
</tr>
<tr>
<td></td>
<td>ZZ = missing data</td>
</tr>
<tr>
<td>DSBJT_CD</td>
<td>type of medical treatment</td>
</tr>
<tr>
<td></td>
<td>01 = internal medicine</td>
</tr>
<tr>
<td></td>
<td>02 = neurology</td>
</tr>
<tr>
<td></td>
<td>03 = psychiatry</td>
</tr>
<tr>
<td></td>
<td>04 = surgery</td>
</tr>
<tr>
<td></td>
<td>05 = orthopedics</td>
</tr>
<tr>
<td></td>
<td>06 = neurosurgery</td>
</tr>
<tr>
<td></td>
<td>07 = cardiothoracic surgery</td>
</tr>
<tr>
<td></td>
<td>08 = plastic surgery</td>
</tr>
<tr>
<td></td>
<td>14 = dermatology</td>
</tr>
<tr>
<td></td>
<td>15 = urology</td>
</tr>
<tr>
<td></td>
<td>16 = radiology</td>
</tr>
<tr>
<td></td>
<td>17 = radiation oncology</td>
</tr>
<tr>
<td></td>
<td>18 = pathology</td>
</tr>
<tr>
<td></td>
<td>19 = laboratory medicine</td>
</tr>
<tr>
<td></td>
<td>20 = tuberculosis</td>
</tr>
<tr>
<td></td>
<td>21 = rehabilitation medicine</td>
</tr>
<tr>
<td>MAIN_SICK</td>
<td>SUB_SICK</td>
</tr>
<tr>
<td>----------------------</td>
<td>---------------------</td>
</tr>
<tr>
<td>main reason of sickness (KCD-6)</td>
<td>minor reason of sickness (KCD-6)</td>
</tr>
<tr>
<td>09 = anesthesia</td>
<td>01</td>
</tr>
<tr>
<td>10 = OB/GYN</td>
<td>11</td>
</tr>
<tr>
<td>11 = pediatrics</td>
<td>12</td>
</tr>
<tr>
<td>12 = ophthalmology</td>
<td>13</td>
</tr>
<tr>
<td>13 = otolaryngology</td>
<td>14</td>
</tr>
<tr>
<td>ZZ = missing data</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Type of medical facility</th>
<th>Healthcare Insurance</th>
<th>Worker’s insurance</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Superior general hospital</td>
<td>22%</td>
<td>45%</td>
<td>15%</td>
</tr>
</tbody>
</table>
### R code used for the final models

```r
# setup - 2013: Identify "diabetes" from MAINT_SICK "E10"-"E14"
setwd("c:/Rdata")
hdata13<-read.csv("NHIS_OPEN_T20_2013.csv")
summary(hdata13)
sick13<-substr(hdata13$MAINT_SICK,1,3)
hdata132<-cbind(hdata13,sick13)
hdata132$sick13_R<-as.factor(hdata132$sick13)
summary(hdata132)
library(plyr)
summary(hdata132)
write.csv(hdata132,"nhdata13.csv")
rm(hdata13)
rm(hdata132)
nhdata13<-read.csv("nhdata13.csv")
summary(nhdata13)

# subset data - 2013: pseudo code to collect all claims from people who have diabetes#
Z1<-nhdata13$MAINT_SICK13 == "diabetes"
Z2<-nhdata13$IDV_ID[Z1]
Z3<-unique(Z2)
Z4<-nhdata13$IDV_ID%in%Z3
Z5<-subset(nhdata13,Z4)
summary(Z5)
write.csv(Z5,"nhdata132.csv")
rm(Z5)
rm(nhdata13)

# subset data - 2013: shrink dataset by patient(IDV_ID) who have diabetes#
nhdata132<-read.csv("nhdata132.csv")
smallnhdata13<-subset(nhdata132, MAIN_SICK13 == "diabetes")
write.csv(smallnhdata13, "nhdata133.csv")
nhdata133<-read.csv("nhdata133.csv")
summary(nhdata133)
```
head(nhdata132)
head(nhdata133)
rm(nhdata132)

#setup - 2008: Identify "diabetes" from MAIN_SICK "E10"-"E14"
setwd("c:/Rdata")
hdata08<-read.csv("NHIS_OPEN_T20_2008.csv")
summary(hdata08)
sick08<-substr(hdata08$MAIN_SICK,1,3)
hdata08<-cbind(hdata08,sick08)
hdata08$sick08_R<-as.factor(hdata08$sick08)
summary(hdata08)
library(plyr)
summary(hdata082)
write.csv(hdata082,"nhdata08.csv")
rm(hdata08)
rm(hdata082)
hdata08<-read.csv("nhdata08.csv")
summary(nhdata08)

#subset data - 2008: pseudo code to collect all claims from people who have diabetes#
Z1<-nhdata08$MAIN_SICK08 == "diabetes"
Z2<-nhdata08$IDV_ID[Z1]
Z3<-unique(Z2)
Z4<-nhdata08$IDV_ID%in%Z3
Z5<-subset(nhdata08,Z4)
summary(Z5)
write.csv(Z5,"nhdata082.csv")
rm(Z5)
rm(nhdata08)

#subset data - 2008: shrink dataset by patient(IDV_ID) who have diabetes#
smallnhdata08<-subset(nhdata08, MAIN_SICK08 == "diabetes")
write.csv(smallnhdata08, "nhdata083.csv")
nhdata083<-read.csv("nhdata083.csv")
summary(nhdata083)
head(nhdata083)
head(nhdata083)
rm(nhdata082)
rm(nhdata083)

#setup - 2003: Identify "diabetes" from MAIN_SICK "E10"-"E14"
setwd("c:/Rdata")
hdata03<-read.csv("NHIS_OPEN_T20_2003.csv")
summary(hdata03)
sick03<-substr(hdata03$MAIN_SICK,1,3)
hdata032<-cbind(hdata03,sick03)
hdata032$sick03_R<-as.factor(hdata032$sick03)
library(plyr)


summary(hdata032)
write.csv(hdata032,"nhdata03.csv")
rm(hdata03)
rm(hdata032)
nhdata03<-read.csv("nhdata03.csv")
summary(nhdata03)

#subset data - 2003: pseudo code to collect all claims from people who have diabetes#
Z1<-nhdata03SMAIN_SICK03 == "diabetes"
Z2<-nhdata03SIDV_ID[Z1]
Z3<-unique(Z2)
Z4<-nhdata03SIDV_ID%in%Z3
Z5<-subset(nhdata03,Z4)
summary(Z5)
write.csv(Z5,"nhdata032.csv")
rm(Z5)
rm(nhdata03)

#subset data - 2003: shrink dataset by patient(IDV_ID) who have diabetes#
nhdata032<-read.csv("nhdata032.csv")
smallnhdata03<-subset(nhdata032, MAIN_SICK03 == "diabetes")
write.csv(smallnhdata03, "nhdata033.csv")
nhdata033<-read.csv("nhdata033.csv")
summary(nhdata033)
head(nhdata032)
head(nhdata033)
rm(nhdata032)

#recoding as factor variables  2013 #
nhdata133<-read.csv("nhdata133.csv")
str(nhdata133)
nhdata133$SEX13_R<-as.factor(nhdata133$SEX)
nhdata133$AGE13_R<-as.factor(nhdata133$AGE_GROUP)
nhdata133$SIDO13_R<-as.factor(nhdata133$SIDO)
nhdata133$FORM_CD13_R<-as.factor(nhdata133$FORM_CD)
nhdata133$DSBJT_CD13_R<-as.factor(nhdata133$DSBJT_CD)
summary(nhdata133)

# Revalue categoric variables 2013#
library(plyr)

nhdata133$SEX13_R2<-mapvalues(nhdata133$SEX13_R, from = c("1", "2"), to = c("Male", "Female"))

```r
nhdata133$FORM_CD13_R2 <- mapvalues(nhdata133$FORM_CD13_R, from = c(2, 3, 6, 7, 8, 9, 10, 11, "ZZ", "-"), to = c("hospital (hospitalization)", "hospital (outpatient)", "maternity clinic (hospitalization)", "public health center (hospitalization)", "public health center (outpatient)", "psychiatric clinic (daytime ward)", "psychiatric clinic (hospitalization)", "psychiatric hospitalization (outpatient)", "missing data", "not applicable")

nhdata133$DSBJT_CD13_R2 <- mapvalues(nhdata133$DSBJT_CD13_R, from = c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, "0"), to = c("internal medicine", "neurology", "psychiatry", "surgery", "orthopedics", "neurosurgery", "cardiothoracic surgery", "plastic surgery", "anesthesia", "OB/GYN", "pediatrics", "ophthalmology", "otolaryngology", "dermatology", "urology", "radiology", "radiation oncology", "pathology", "laboratory medicine", "tuberculosis", "rehabilitation medicine", "nuclear medicine", "family medicine", "emergency medicine", "occupational & environmental", "preventive medicine", "NA"))
```

```r
summary(nhdata133)
write.csv(nhdata133, "nhdata133.csv")
```

# recoding as factor variables 2008 #
```

```r
library(plyr)

```

```r
# Revalue categoric variables 2008#
```

```r
library(plyr)
```

```r
```

```r
```

summary(nhdata083)
write.csv(nhdata083, "nhdata083.csv")

# recoding as factor variables 2003#

# recoding as factor variables 2003#

library(plyr)

nhdata033$SEX03_R2 <- mapvalues(nhdata033$SEX03_R, from = c("1", "2"), to = c("Male", "Female"))

library(ggplot2)
figure1<-ggplot(nhdata133, aes(AGE13_R2, fill=SEX13_R2))+geom_bar(position = "dodge")+scale_x_discrete("Age Group")+scale_y_continuous("Number of Diabetes Patients")+theme_grey(base_size=27)+theme(axis.text.x=element_text(angle = -60, hjust = 0))+theme(legend.position="bottom") +theme(legend.key=element_blank(), legend.title=element_blank(), legend.box="horizontal")
figure1

# 2013: Figure 2 #
figure2<-qplot(AGE13_R2, nhdata133, geom="bar", weight=EDEC_TRAMT, ylab="Total Costs", xlab="Age Group", fill=FORM_CD13_R2)+theme_grey(base_size=27)+theme(axis.text.x=element_text(angle = -60, hjust = 0))+theme(legend.position="bottom") +theme(legend.key=element_blank(), legend.title=element_blank(), legend.box="horizontal")
figure2

# cost 2013: Figure 3#
cdata133 <- dplyr::ddply(nhdata133, c("IDV_ID","AGE13_R2","FORM_CD13_R2"), summarise, N = length(EDEC_TRAMT), totals = sum(EDEC_TRAMT), Mean=mean(EDEC_TRAMT), max = max(EDEC_TRAMT), nhis=mean(EDEC_JBRDN_AMT))
head(cdata133)
cdata134 <- dplyr::ddply(cdata133, c("AGE13_R2","FORM_CD13_R2"), summarise, N = mean(Mean), totals = sum(mean), TotalPayment=mean(mean), max = max(mean), NHISPayment=mean(nhis))
cdata134

figure3<-ggplot(cdata134, aes(AGE13_R2, TotalPayment, fill=FORM_CD13_R2))+geom_bar(stat="identity", position="dodge")+labs(x="Age Group", y="Total Payment per treated Patient(won)")+theme_grey(base_size=27)+theme(axis.text.x=element_text(angle = -60, hjust = 0))+theme(legend.position="bottom") +theme(legend.key=element_blank(), legend.title=element_blank(), legend.box="horizontal")
figure3

figure3<-ggplot(cdata134, aes(AGE13_R2, TotalPayment, fill=FORM_CD13_R2))+geom_bar(stat="identity", position="dodge")+scale_y_sqrt()+labs(x="Age Group", y="Total Payment per treated Patient(won)")+theme_grey(base_size=27)+theme(axis.text.x=element_text(angle = -60, hjust = 0))+theme(legend.position="bottom") +theme(legend.key=element_blank(), legend.title=element_blank(), legend.box="horizontal")
figure3

# cost 2013: Figure 4#
hndata133<-read.csv("nhdata133.csv")
library(plyr)
cdata131 <- dplyr::ddply(nhdata133, c("IDV_ID","AGE13_R2"), summarise, N = length(EDEC_TRAMT), totals = sum(EDEC_TRAMT), Mean=mean(EDEC_TRAMT), max = max(EDEC_TRAMT), nhis=mean(EDEC_JBRDN_AMT))
head(cdata131)
cdata132 <- dplyr::ddply(cdata131, c("AGE13_R2"), summarise, N = length(Mean),TotalPayment=mean(Mean), max = max(Mean), NHISPayment=mean(nhis))
cdata132
```r
# 2013: Figure 5 #
table1 <- xtabs(~DSBJT_CD13_R2, data=nhdata133)
table1p <- 100*prop.table(table1)
table1p
library(googleVis)
dat <- data.frame(DSBJT=c("Internal medicine", "Family medicine", "Surgery", "Orthopedics", "Pediatrics", "Others"), percet=c(90.12, 4.41, 2.08, 0.9, 0.54, 1.95))
figure5 <- gvisPieChart(dat, options=list(width=700, height=700, slices="{0: {offset: 0.2}, 1: {offset: 0.2}, 2: {offset: 0.2}}", title='Figure5. Type of Medical Treatment for Diabetes Claims in 2013', legend="{ position: 'right', textStyle: {fontSize: 12}}", chartArea="{left:25,top:30}"), colors="[blue', 'orange', 'yellow', 'skyblue', 'purple', 'lightgreen']", pieSliceText='label', pieHole=0.5), chartid="figure1")
plot(figure5)
# cost 2013: Figure 6#
cdata135 <- ddply(nhdata133, c("IDV_ID","DSBJT_CD13_R2"), summarise, N = length(EDEC_TRAMT), totals = sum(EDEC_TRAMT), mean=mean(EDEC_TRAMT), max = max(EDEC_TRAMT), nhis=mean(EDEC_JBRDN_AMT))
head(cdata135)
cdata136 <- ddply(cdata135, c("DSBJT_CD13_R2"), summarise, N = length(mean), totals = sum(mean), TotalPayment=mean(mean), max = max(mean), NHISPayment=mean(nhis))
cdata136$DSBJT_CD13_R2 <- factor(cdata136$DSBJT_CD13_R2, levels = cdata136$DSBJT_CD13_R2[order(cdata136$TotalPayment)])
cdata136$DSBJT_CD13_R2
figure6<-ggplot(cdata136, aes(DSBJT_CD13_R2)) + geom_bar(aes(y = TotalPayment, fill="TotalPayment"), stat="identity") + geom_bar(aes(y = NHISPayment, group = 1, color = "NHISPayment"), stat="identity") +scale_colour_manual("", values=c("NHISPayment" = "grey", "TotalPayment" = "grey"))+ scale_fill_manual("", values="grey")+labs(x="Type of Medical Treatment", y="Total/NHIS Payment per Patient(won)")+theme_grey(base_size=25)+theme(axis.text.x=element_text(angle = -60, hjust = 0))+theme(legend.position="right")
```

```r
# cost trends: Figure 7 #
data131 <- ddply(nhdata133, c("IDV_ID","AGE13_R2"), summarise, N = length(EDEC_TRAMT), totals = sum(EDEC_TRAMT), mean=mean(EDEC_TRAMT), max = max(EDEC_TRAMT), nhis=mean(EDEC_JBRDN_AMT))
head(cdata131)
cdata132 <- ddply(cdata131, c("AGE13_R2"), summarise, N = length(mean), totals = sum(mean), TotalPayment=mean(mean), max = max(mean), NHISPayment=mean(nhis))
cdata132
```
nhdata083 <- read.csv("nhdata083.csv")
cdata081 <- ddply(nhdata083, c("IDV_ID","AGE08_R2"), summarise, N = length(EDEC_TRAMT),
totals = sum(EDEC_TRAMT), mean= mean(EDEC_TRAMT), max = max(EDEC_TRAMT),
his=mean(EDEC_JBRDN_AMT))
head(cdata081)
cdata082 <- ddply(cdata081, c("AGE08_R2"), summarise, N = length(mean),
totals = sum(mean), TotalPayment=mean(mean), max = max(mean), NHISPayment=mean(nhis))
cdata082

nhdata033<- read.csv("nhdata033.csv")
cdata031 <- ddply(nhdata033, c("IDV_ID","AGE03_R2"), summarise, N = length(EDEC_TRAMT),
totals = sum(EDEC_TRAMT), mean= mean(EDEC_TRAMT), max = max(EDEC_TRAMT),
his=mean(EDEC_JBRDN_AMT))
head(cdata031)
cdata032 <- ddply(cdata031, c("AGE03_R2"), summarise, N = length(mean),
totals = sum(mean), TotalPayment=mean(mean), max = max(mean), NHISPayment=mean(nhis))
cdata032

trend1<- read.csv("trend13.csv")
trend1
y1<-trend1$X2003
y2<-trend1$X2008
y3<-trend1$X2013
plot(trend1$AGE, y2, type="b", pch=19,col="red",xlab="Age Group", ylab="Costs of Diabetes Treatments by Age (won)"
lines(trend1$AGE, y3, pch=18, col="blue", type="b")
lines(trend1$AGE, y1, pch=17, col="darkgreen", type="b")
legend("center",legend=c("2003","2008", "2013"),col=c("darkgreen", "red","blue"),bg="white",lwd=2)

# cost trends: map Figure 8, 9, 10#
cdata137 <- ddply(nhdata133, c("IDV_ID","SIDO13_R"), summarise, N = length(EDEC_TRAMT),
totals = sum(EDEC_TRAMT), mean= mean(EDEC_TRAMT), max = max(EDEC_TRAMT),
his=mean(EDEC_JBRDN_AMT))
head(cdata137)
cdata138 <- ddply(cdata137, c("SIDO13_R"), summarise, N = length(mean), totals = sum(mean),
TotalPayment=mean(mean), max = max(mean), NHISPayment=mean(nhis))
cdata138

library(sp)
library(colorspace)
kmap<-readRDS("KOR_adm1.rds")
plot(kmap)
kmap@data
sample <- data.frame(ID_1 = c(16,1,5,11,7,4,17,15,8,6,2,3,13,14,9,10,12L), DIA =
c(36926,44336,43199,33029,33029,33029,33029,33029,33029,33029,33029,33029,33029,33029,33029,33029)
final <- merge(x =kmap@data, y = sample, by = "ID_1", all.y = TRUE)
kmap@data <- data.frame(kmap@data, sample[match(kmap@data[,"ID_1"], sample[,]"ID_1")],)
final[ order(-finalSDIA),c("ID_1","DIA") ]
kmap$DIA <- final$DIA
clds <- c('#FCFC8E', '#F8F6C5', '#EAE44A', '#E5DF27',
            '#DFD81A', '#CDC718', '#BCB616', '#AAA514',
            '#999412', '#878310', '#75720E', '#64610C',
            '#52500A', '#413F08', '#2F2E06', '#1E1D04', '#0C0C01')
spplot(kmap, zcol = 'DIA', main = 'Regional Variation in Incidence of Diabetes (2013)', col.regions = clds)

cdata087 <- ddply(nhdata083, c('IDV_ID', 'SIDO08_R'), summarise,
            N = length (EDEC_TRAMT), totals = sum(EDEC_TRAMT),
            mean = mean(EDEC_TRAMT), max = max(EDEC_TRAMT),
            nhis = mean(EDEC_JBRDN_AMT))
head(cdata087)
cdata087 <- ddply(cdata087, c('SIDO08_R'), summarise,
            N = length(mean), totals = sum(mean),
            TotalPayment = mean(mean), max = max(mean),
            NHISPayment = mean(nhis))
cdata087

cdata037 <- ddply(nhdata033, c('IDV_ID', 'SIDO03_R'), summarise,
            N = length(EDEC_TRAMT), totals = sum(EDEC_TRAMT),
            mean = mean(EDEC_TRAMT), max = max(EDEC_TRAMT),
            nhis = mean(EDEC_JBRDN_AMT))
head(cdata037)
cdata038 <- ddply(cdata037, c('SIDO03_R'), summarise,
            N = length(mean), totals = sum(mean),
            TotalPayment = mean(mean), max = max(mean),
            NHISPayment = mean(nhis))
cdata038

kmap <- readRDS("KOR_adm1.rds")
plot(kmap)
kmap@data
sample <- data.frame(ID_1 = c(16, 1, 5, 11, 7, 4, 17, 15, 8, 6, 2, 3, 13, 14, 9, 10, 12L),
            DIA = c(42573, 34162, 28130, 31377, 29957, 34016, 39306, 39011, 34797, 31972,
            35113, 30102, 28604, 31547, 3126, 29201L))
final <- merge(x = kmap@data, y = sample, by = 'ID_1', all.y = TRUE)
kmap@data <- data.frame(kmap@data, sample[match(kmap@data[, 'ID_1'], sample[, 'ID_1'])],)
final[ order(-final$DIA), c('ID_1', 'DIA') ]
kmap$DIA <- final$DIA
clrs <- c("FCFCE8", "F8F6C5", "EAE44A", "E5DF27", "DFD81A", "CDC718", "BCB616", "AA5414", "999412", "78310", "75720E", "64610C", "52500A", "413F08", "2F2E06", "1E1D04", "0C0C01")

spplot(kmap, zcol = "DIA", main = "Regional Variation in Incidence of Diabetes (2003)", col.regions = clrs)

# Rescale TotalPayment.
crs$dataset["R10_TotalPayment"] <- crs$dataset["TotalPayment"]

# Take a log10 transform of the variable - treat -Inf as NA.
if (building)
  crs$dataset["R10_TotalPayment"] <- log10(crs$dataset["TotalPayment"])
crs$dataset[crs$dataset["R10_TotalPayment"] == -Inf & ! is.na(crs$dataset["R10_TotalPayment"])), "R10_TotalPayment"] <- NA

# When scoring transform using the training data parameters.
if (scoring)
  crs$dataset["R10_TotalPayment"] <- log10(crs$dataset["TotalPayment"])
crs$dataset[crs$dataset["R10_TotalPayment"] == -Inf & ! is.na(crs$dataset["R10_TotalPayment"])), "R10_TotalPayment"] <- NA

# Note the user selections.
# The following variable selections have been noted.
crs$input <- c("AGE_R", "Facility", "Sex", "Treatment", "Location", "Hospitalization", "Medication")
crs$numeric <- c("Hospitalization", "Medication")
crs$categoric <- c("AGE_R", "Facility", "Sex", "Treatment", "Location")
crs$target <- "R10_TotalPayment"
crs$risk <- NULL
crs$ident <- "IDV_ID"
crs$ignore <- c("TotalPayment", "Age")
crs$weights <- NULL

# Random Forest
# The 'randomForest' package provides the 'randomForest' function.
library(randomForest, quietly=TRUE)

# Build the Random Forest model.
set.seed(crv$seed)
crs$rf <- randomForest::randomForest(R10_TotalPayment ~ .,
  data=crs$dataset[crs$sample, c(crs$input, crs$target)],
  ntree=500,
  mtry=2,
  importance=TRUE,
na.action=randomForest::na.roughfix, replace=FALSE)

# Generate textual output of 'Random Forest' model.
crs$rf

# List the importance of the variables.n <- round(randomForest::importance(crs$rf), 2)
rn[order(rn[,1], decreasing=TRUE),]

# Time taken: 9.48 mins

# Plot the relative importance of the variables.
randomForest::varImpPlot(crs$rf, main="")
title(main="Variable Importance Random Forest cdata140.csv",
     sub=paste("Rattle", format(Sys.time(), "%Y-%b-%d %H:%M:%S"), Sys.info()["user"])))

# Plot the error rate against the number of trees.
plot(crs$rf, main="")
legend("topright", c(""), text.col=1:6, lty=1:3, col=1:3)
title(main="Error Rates Random Forest cdata140.csv",
     sub=paste("Rattle", format(Sys.time(), "%Y-%b-%d %H:%M:%S"), Sys.info()["user"])))

# Plot the OOB ROC curve.
library(verification)
aucc <- verification::roc.area(as.integer(as.factor(crs$dataset[crs$sample, crs$target]) - 1, crs$rf$votes[,2])$A
verification::roc.plot(as.integer(as.factor(crs$dataset[crs$sample, crs$target]) - 1, crs$rf$votes[,2], main="")
legend("bottomright", bty="n",
     sprintf("Area Under the Curve (AUC) = %1.3f", aucc))
title(main="OOB ROC Curve Random Forest cdata140.csv",
     sub=paste("Rattle", format(Sys.time(), "%Y-%b-%d %H:%M:%S"), Sys.info()["user"])))

# Evaluate model performance.
library(ggplot2)
# Generate a risk chart.
library(ggplot2)
# Rattle provides evaluateRisk() and riskchart().
crs$pr <- predict(crs$rf, newdata=na.omit(crs$dataset[crs$validate, c(crs$input, crs$target)]))
crs$eval <- evaluateRisk(crs$pr, na.omit(crs$dataset[crs$validate, c(crs$input, crs$target)])$R10_TotalPayment)
print(riskchart(crs$pr,
       na.omit(crs$dataset[crs$validate, c(crs$input, crs$target)])$R10_TotalPayment,
       title="Performance Chart Random Forest cdata140.csv [validate] ", show.lift=FALSE,
       show.precision=FALSE, legend.horiz=FALSE))

# Evaluate model performance.
# RF: Generate a Predicted v Observed plot for rf model on cdata140.csv [validate].
crs$pr <- predict(crs$rf, newdata=na.omit(crs$dataset[crs$validate, c(crs$input, crs$target)]))
# Record rows omitted from predict command.
omitted <- attr(na.omit(crs$dataset[crs$validate, c(crs$input, crs$target)]), "na.action")
# Obtain the observed output for the dataset.
obs <- subset(crs$dataset[crs$validate,][-omitted,], select=crs$target)
# Handle in case categoric target treated as numeric.
obs.rownames <- rownames(obs)
obs <- as.numeric(obs[[1]])
obs <- data.frame(R10_TotalPayment=obs)
rownames(obs) <- obs.rownames
# Combine the observed values with the predicted.
fitpoints <- na.omit(cbind(obs, Predicted=crs$pr))
# Obtain the pseudo R2 - a correlation.
fitcorr <- format(cor(fitpoints[,1], fitpoints[,2])^2, digits=4)
# Plot settings for the true points and best fit.
op <- par(c(lty="solid", col="blue"))
# Display the observed (X) versus predicted (Y) points.
plot(fitpoints[,1], fitpoints[,2], asp=1, xlab="R10_TotalPayment", ylab="Predicted")
# Generate a simple linear fit between predicted and observed.
prline <- lm(fitpoints[,2] ~ fitpoints[,1])
# Add the linear fit to the plot.
abline(prline)
# Add a diagonal representing perfect correlation.
par(c(lty="dashed", col="black"))
abline(0, 1)
# Include a pseudo R-square on the plot
legend("bottomright", sprintf(" Pseudo R-square=%s ", fitcorr), bty="n")
# Add a title and grid to the plot.
title(main="Predicted vs. Observed Random Forest Model cdata140.csv [validate]",
sub=paste("Rattle", format(Sys.time(), "%Y-%m-%d %H:%M:%S"), Sys.info()["user"]))
grid()