Analytics of Patients and Consumers Survey

Healthcare Data Analytics and Data Mining

Group 1

Heyuan Gao
Zijing Hu
Jinman Rong
Ziyu Tang
Pengcheng Xu
Haina Yan

Introduction

Patients and consumers surveys are becoming an important source of information about the performance and responsiveness of healthcare systems. It can help evaluate healthcare systems and improve the quality of healthcare by being more patient centered.

In this report, we use the Public Use File (PUF) of 2016 US Medicare Current Beneficiary Survey (MCBS) to analyze several interesting topics:

- Racial disparity in ability to pay for care
- Gender differentials in healthcare utilization
- Education and health
- Obesity and anxiety/depression
- The role of gender in the severity of the obesity-depression causal relationship
- Loneliness and health
- Loneliness and risk of depression

Racial Disparity in Ability to Pay for Care

In this question, we examine if Medicare beneficiaries older than 65 belonging to different racial groups show distinctions in financial ability. In particular, we would like to compare the financial ability of Non-Hispanic white with that of Non-Hispanic black. We used the 'ACC_HCDELAY': 'Last year ever delay in care due to cost' as a proxy variable for financial ability indication. For this variable, the value 1 = Yes means the person has delayed care for money reasons. We also filtered out value = R or D for this variable, since these two values mean interviewees are not willing to reveal their status with regards to this variable and are equivalent to null values.

As a result, we construct the following 2*2 table:

Table 1 Contingency Table of Racial Disparity and Pay Ability

	White	Black
Financial difficulty	462	77
Non-Financial difficulty	7,731	772
odds ratio		0.5591885
p-value		0.0001741

In the columns are races to be investigated, and the rows indicate if the beneficiary delays in care due to cost last year. As we can see from the table, of the two races we investigated, 539 people reported delayed payment for healthcare service last year, indicating they are having financial difficulties. That number, converting to ratio, is 5.96%. Of the people who reported delayed payment, white people account for 14.29%, whereas, black people account for 85.71%, presenting astonishing racial disparity in terms of payment ability for healthcare services.

Moreover, people with financial difficulties account for only 5.98% among the white, whereas, people with the same difficulties account for 9.97% among the black, twice as much as the proportion among white people.

We then conducted a Fisher test, and results of the test is shown below:

Odds ratio in this case equals to 0.6. It is interpreted as following: the odds of delaying care last year owing to cost if the beneficiary is white is 0.6 compared to if the beneficiary is black with a 95% confidence interval [0.46, 0.78]. The hypothesized odds ratio equals 1. Consequently, p-value equals 0.0001, significantly smaller than 1%. Thus we reach the conclusion that there is racial disparity with regard to financial difficulties and the difference is statistically significant.

Gender Differentials in Healthcare Utilization

On average, women utilize more health services than men. One of the reasons could be that women need more care during reproductive ages which results in the gender differences. However, we suppose that such differences in health seeking behavior sustains as they go old as well. To examine that women naturally have better health seeking behavior than men, we conduct a one-sided t-test to see whether the health utilization of female is significantly higher than that of male in the elder age.

Our null hypothesis and alternative hypothesis are shown as below:

H0: Females' average health utilization is no greater than that of males
Ha: Females' average health utilization is greater than that of males

For the data processing, we first filter the data with ADM_H_MEDSTA to be 1 to insure only individuals with age over 65 are included for analysis. Then, we create a new continuous variable, healcare utilization, according to the midpoint of all count ranges of ADM_H_PHYEVT. With the new continuous variable, we further calculate the weighted average of number of doctor visits for males and females subjects separately. Finally, we conduct the t-test to investigate whether the health utilization differences in gender is significant.

According to the result of t-test, we can reject the null hypothesis. There is significant gender difference in the healthcare utilization, and females' average health utilization is significantly greater than that of males at 1% level. (t(10074) = 2.8938, p = 0.001907)

Furthermore, the weighted average of number of doctor visits for females (5.256449) is greater than that of males (4.891393).

Table 2 t-test of Gender Differentials in Healthcare Utilization

	Female Utilization	Male Utilization
Weighted average	5.256449	4.891393
Observations	5,970	4,659
df		10074
t Stat		2.8938
p-value		0.001907

To sum up, we agree that men are in fact lazier or ignorant in terms of health seeking behavior, and this fact is not necessarily because women are assuming the burden of childbearing and caring when they are in reproductive ages because the pattern continues onto older age as well.

Relationship Between Education and Health

In this case, we tried to test the hypothesis that education makes individuals healthier. This hypothesis is raised from the health econ textbook and the author indicates the reason may be highly educated people have better lifestyle and more informed life and health related decisions. To measure health conditions, we used a naïve indicator – obesity (the variable HLT_BMI_CAT). We considered people who were obese or had extreme or high-risk obesity (HLT_BMI_CAT = 4 and 5) as worse health condition and others as healthy ones. In terms of education, we treated patients who had higher education level than high school as highly educated (DEM_EDU = 3) and others were less educated. Noted that we ignored the null value in both variables (such as 'D', 'N' and 'R' in DEM_EDU).

In order to see the difference in health conditions between highly educated group and less educated group, we created a following contingency table:

Table 3 Contingency Table of Health and Depression

	Healthy	Obesity
Highly Educated	4014	1591
Less Educated	4599	2208

Accordingly, we applied Fisher Exact Test to see if the difference is significant. Based on the table above, the odds ratio is 1.211253 and the P value is 1.097e-06. The difference between the two education groups is significant. Since the odds ratio is higher than 1, it proves our hypothesis that highly educated people tend to have better health conditions.

Obesity and Depression

Some medical literature claims the increased risk of depression and anxiety disorder among obese individuals. Additionally, some believe that depression result in obesity and eating disorder. Thus, we will determine the relationship between obesity and depression in this part.

In this dataset, we use HLT_BMI_CAT to determine whether person obese, if the value is larger than 3(HLT_BMI_CAT = 4 and 5), it means person has obesity. Otherwise, if values are smaller or equal to 3, it means person doesn't have obesity. As for depression, we use HLT_OCDEPRSS to determine whether people with depression, with value 1 means Yes, the person has been diagnosed with depression.

To see the depression rate in the two healthy and obese groups and if the difference in the depression rate is in fact statistically different between the two groups, we built a table, shown as follows:

Table 4 Contingency Table of Health and Depression

	Obesity	Healthy
Depression	1366	2039
No-Depression	2446	6593

From above, we can calculate depression rate for healthy groups are 2039 / (2039 + 6593) = 23.62%. Depression rate for obesity groups are 1366 / (1366+2446) = 35.83%

Then we conducted the Fisher's exact test to see whether there is the relationship between depression and obesity. According to the exhibit below, the odds ratio is 1.805713. The p value is 2.2e-16, which is significant. Additionally, 95% confidence interval is 1.66 to 1.96, which doesn't contain value 1. Thus, we can reject the null hypothesis. In conclusion, the difference in the depression rate is statistically different between the two groups.

Gender in the severity of the obesity-depression causal relationship

In question 5, our task is to examine the role of gender in the obesity-depression causal relationship. Consider that for each gender we can generate a 2×2 cross tabulation to show the number of people with different health conditions, instead of comparing the results of Fisher exact test of each cross tabulation, we used Cochran-Mantel-Haenszel Chi-Squared Test (CMH Test) to test whether gender has a statistically significant impact on the relationship between obesity and depression.

Table 5 Cross Tabulation of Gender, Obesity and Depression

Obesity	Depression	Male	Female
Yes	Yes	491	875
	No	1189	1257
	Percentage	29.2%	41.0%
No	Yes	779	1260
	No	3236	3357
	Percentage	19.4%	27.2%

After removing rows containing meaningless values and transforming column "HLT_BMI_CAT" into a binary variable, we have three nominal variables: gender (male or female), obesity (yes or no) and depression (yes or no). The table above shows the number of persons with different diseases. There is a smaller proportion of depression rate in male; we want to know whether this difference is significant.

We set the null hypothesis that the proportion of people with depression is the same in male and female (true common odds ratio is equal to 1) and got a Mantel-Haenszel chi-squared of 192.26 with p-value less than 2.2e-16. Thus, we made the statement that the differences caused by gender are statistically significant.

Loneliness and health

In this part, we test how the seniors who live alone without a family or any effective social relationships perceive their own health. In particular, we compare the perception of own health of older people who live with family with those who live alone. Here, we use "DEM_MARSTRA" to determine the loneliness. The value is equal to 1 means the person lives with family and not feel alone while the value is equal to 2,3,4 means the person lives alone. The other variable is "HLT_GENHELTH", which means the responders perception of the general health compared to others at the same age. The value is equal to 1 or 3 when the person feels happy and healthy while the value is equal to 4 or 5 when the person feels their health is worse than others. One thing needs to mention is that we do use the variable "ADM_H_MEDSTA=1" to filter the data in order to ensure that only people aged over 65 are studied in this part.

To examine the significance of the association (contingency) between those two classifications, we built a contingency table, shown as follows:

Table 6 Contingency Table of Health and Loneliness

	With Family	Living Alone
Poor Fair Health	895	1049
Good Health	4590	4033

Then we conducted the Fisher's exact test to demonstrate whether there is association between health and loneliness (odds ratio is equal to 1). In the result, odds ratio is 0.75 and the p-value is significant, and the confidence interval doesn't contain 1. Therefore, we reject the null hypothesis and conclude that the odds of people with a poor fair health if they are living with family is 0.75 compared to if people with a good health but are living alone. We get a p-value that is smaller than 0.05, which means that there is a significant association between loneliness and health.

Loneliness and risk of depression

This question is an extension of the previous one. In this part, we discuss the relationship between loneliness and risk of depression. We first use the variable "ADM_H_MEDSTA=1" to filter the data in order to ensure that only people aged over 65 are studied in this part. Then we choose another variable to represent depression, which is HLT_OCDEPRSS. If the value is equal to 1, it means the person has been diagnosed with depression.

However, we use two methods to define people without depression:

Firstly, if HLT_OCDEPRSS is equal to 2, it means the person has not been diagnosed with depression. The contingency table is built as follows:

Table 7 Contingency Table of Depression and Loneliness (Method 1)

	With Family	Living Alone
Depression	992	1255
Not Depression	4513	3839

Secondly, if HLT_OCDEPRSS is not equal to 1, it means the person does not officially declare that he has depression. In this case, we labelled all of them as not depression. The contingency table is built as follows:

Table 8 Contingency Table of Depression and Loneliness (Method 2)

	With Family	Living Alone
Depression	992	1255
Not Depression	4518	3854

We conducted the Fisher's exact test for each of the situation to demonstrate whether there is association between depression and loneliness (odds ratio is equal to 1). As a result, odds ratio is all equal to 0.67. Therefore, we reject the null hypothesis and conclude that the odds of people with depression if they are living with family is 0.67 compared to the people without depression but are living alone. We get a p-value that is smaller than 0.05 and the confidence interval doesn't contain 1, which mean that there is a significant association between loneliness and depression.

Summary

By analyzing 2016 US Medicare Current Beneficiary Survey to analyze, we test the racial disparity in ability to pay for care, gender differentials in healthcare utilization, relationship between education and health, we also identify how obesity relates to depression, the role gender plays in obesity-depression causal relationship, and interpret the association between loneliness and risk of getting physical and mental diseases.

From our findings, non-Hispanic white patients significantly have better ability to pay for the service compared with non-Hispanic black patients. A significant difference was identified between male and female average hospital utilizations. Highly educated individuals had lower risk of obesity compared with less educated individuals, and this relationship can be even more significant when we consider gender factor. There is a signification correlation between obesity and depression where obese people being more at risk of getting depressed than the non-obese people. We also find that people who live by themselves had higher risk of mental health problems, such as depression, relative to those who live with their spouse, children or other family members. Also, those who live alone state that their perception of the general health is worse compared to others at the same age. Thus, caring for your body shape and companion by others can be helpful for maintaining people's physical and mental health.

Appendix

~~~~
```{r}
library(dplyr)
puf <- read.csv("./puf2016.csv")
# import survey data.
```


```{r}
puf <- puf %>% filter(ADM_H_MEDSTA==1)

puf.w <- puf %>% filter(DEM_RACE==1)
puf.b <- puf %>% filter(DEM_RACE==2)

puf.w <- puf.w %>% filter(ACC_HCDELAY %in% c(1,2))
puf.b <- puf.b %>% filter(ACC_HCDELAY %in% c(1,2))

puf.w1 <- puf.w %>% filter(ACC_HCDELAY==1)
puf.w2 <- puf.w %>% filter(ACC_HCDELAY==2)
puf.b1 <- puf.b %>% filter(ACC_HCDELAY==1)
puf.b2 <- puf.b %>% filter(ACC_HCDELAY==2)

# filter white/delayed to puf.w1; white/not-delayed to puf.w2; black/delayed to puf.b1; black/not-delayed to puf.b2.
```

```{r}
d=matrix(c(nrow(puf.w1), nrow(puf.w2), nrow(puf.b1), nrow(puf.b2)), 2,2,
          dimnames = list(c('Delayed', 'No-Delay'), c('White', 'Black')))
# Construct a 2*2 matrix.
d
```

```{r}
fisher.test(d)
```
```{r loadlib, echo=T, results='hide', message=F, warning=F}
rm(list=ls())
library(data.table)
library(sandwich)
library(tidyverse)
library(lmtest)
library(ggplot2)
library(knitr)
library(psych)
library(dplyr)
```


## Load data


```{r}
puf2016 = read_csv("puf2016.csv")
```

### Q2) Gender differentials in healthcare utilization

```{r}
q2 = puf2016[puf2016$ADM_H_MEDSTA==1,]
q2 = q2[,c("ADM_H_PHYEVT","DEM_SEX")]
q2 = q2 %>% mutate(ADM_H_PHYEVT = case_when(ADM_H_PHYEVT==1~3,
                                            ADM_H_PHYEVT==2~8,
                                            ADM_H_PHYEVT==3~13,
                                            ADM_H_PHYEVT==4~18,
                                            ADM_H_PHYEVT==5~23,
                                            TRUE~0))
names(q2) = c("Health_Utilization","Gender")
female_utilization = q2[q2$Gender==2,]
male_utilization   = q2[q2$Gender==1,]
unique(q2$Health_Utilization)
```


```{r}
female_average = mean(female_utilization$Health_Utilization)
~~~~

~~~
male_average = mean(male_utilization$Health_Utilization)
female_average
male_average
```



```{r}
#t.test(female_utilization$Health_Utilization,male_utilization$Health_Utilization)
t.test(female_utilization$Health_Utilization,male_utilization$Health_Utilization,alternative = "greater")
```

## import packages and initialize datasets
```{r message=FALSE, warning=FALSE}
library(tidyverse)
library(reshape2)
library(tidyverse)

puf2016 <- read.csv('./MCBSpuf2016/puf2016.csv')
```

## Fisher's Exact Test for HLT_BMI_CAT and DEM_EDU

```{r}
puf2016.use <- puf2016[c('PUF_ID', 'HLT_BMI_CAT', 'DEM_EDU')]

puf2016.use <- within(puf2016.use, {
    HLT_BMI_CAT[HLT_BMI_CAT>=4] <- 'Obesity'
    HLT_BMI_CAT[HLT_BMI_CAT<4] <- 'Healthy'
    DEM_EDU[DEM_EDU %in% c('D', 'N', 'R')] <- NA
    DEM_EDU <- as.integer(DEM_EDU)
    DEM_EDU[DEM_EDU==3] <- 'Highly Educated'
    DEM_EDU[DEM_EDU<3] <- 'Less Educated'}
)

ftable <- dcast(na.omit(puf2016.use), HLT_BMI_CAT ~ DEM_EDU, length)
ftable <- data.frame(ftable[2:3], row.names = ftable$HLT_BMI_CAT)
ftable
fisher.test(ftable)
```
---
title: "Assignment_1.4_Obesity depression"
author: "Haina Yan"
date: "3/28/2020"
output: html_document
---

```{r}
library(dplyr)
puf <- read.csv("~/Desktop/puf2016.csv")
# import survey data.
```




```{r}
puf.4<-puf[c("PUF_ID","HLT_OCDEPRSS","HLT_BMI_CAT")]

puf.o <- puf.4 %>% filter(HLT_BMI_CAT>=4)
puf.n <- puf.4 %>% filter(HLT_BMI_CAT<=3)

puf.o1 <- puf.o %>% filter(HLT_OCDEPRSS==1)
puf.o2 <- puf.o %>% filter(HLT_OCDEPRSS==2)
puf.n1 <- puf.n %>% filter(HLT_OCDEPRSS==1)
puf.n2 <- puf.n %>% filter(HLT_OCDEPRSS==2)

# filter Obesity/Depression to puf.o1; Obesity/No-Depression to puf.o2;
# No-Obesity/Depression to puf.n1;    No-Obesity/No-Depression to puf.n2

d4=matrix(c(nrow(puf.o1), nrow(puf.o2), nrow(puf.n1), nrow(puf.n2)), 2,2,
              dimnames = list(c('Depression', 'No-Depression'), c('Obesity', 'NO-Obesity')))

# Construct a 2*2 matrix.
d4
fisher.test(d4)
```
```{r}
# Import required package
library(data.table)
~~~

~~~
```

```{r}
# Load .csv data file
dt <- fread('puf2016.csv')

# Drop rows with meanless values
dt <- subset(dt, DEM_SEX %in% c(1,2))
dt <- subset(dt, HLT_BMI_CAT %in% 1:5)
dt <- subset(dt, HLT_OCDEPRSS %in% c(1,2))

# Recode feature "HLT_BMI_CAT"
set(dt, i = which(dt$HLT_BMI_CAT < 4), 'HLT_BMI_CAT', 2)
set(dt, i = which(dt$HLT_BMI_CAT >= 4), 'HLT_BMI_CAT', 1)

# Set columns for further analysis
cols <- c('DEM_SEX', 'HLT_BMI_CAT', 'HLT_OCDEPRSS')

# Segment data table into two subsets by gender
dt.m <- subset(dt, DEM_SEX == 1)[, cols, with = F]
dt.f <- subset(dt, DEM_SEX == 2)[, cols, with = F]
```

```{r}
crosstab <- function(d) {
   # Input: a data.table contains column HLT_BMI_CAT and column HLT_OCDEPRSS
   # Output: a matrix format cross tabulation of HLT_BMI_CAT and HLT_OCDEPRSS
   mat <- as.matrix(dcast(d, HLT_BMI_CAT ~ HLT_OCDEPRSS, length)[,2:3])
   colnames(mat) <- c('depression', 'no depression')
   rownames(mat) <- c('obesity', 'no obesity')
   return(mat)
}

ma.m = crosstab(dt.m)
ma.f = crosstab(dt.f)
```

```{r}
# Construct a 3-D cross tabulation
segment <- array(c(ma.m, ma.f),
                         dim = c(2, 2, 2),
                         dimnames = list(
                              Obesity = c("Yes", "No"),
                              Depression = c("Yes", "No"),
                              Sex = c("M", "F")))
segment
```

```{r}
# Cochran–Mantel–Haenszel test
mantelhaen.test(segment)
```

```{r, include=FALSE}
rm(list = ls())
gc()
```

```{r, message=FALSE}
library(data.table)
library(dplyr)
```

```{r}
# import data
MCBS = read.csv("/Users/jemma/Desktop/Health analysis and data mining/2020 spring/session 1/MCBSpuf2016/puf2016.csv")
```

### Q6) Loneliness and health
```{r}
# filter data
PoorFairHealth_t <- MCBS %>% filter(ADM_H_MEDSTA == 1) %>% filter(HLT_GENHELTH %in% c(4,5))
GoodHealth_t <- MCBS %>% filter(ADM_H_MEDSTA == 1) %>% filter(HLT_GENHELTH %in% c(1,2,3))
PoorFairHealth_WithFamily <- PoorFairHealth_t %>% filter(DEM_MARSTA == 1) %>% nrow()
PoorFairHealth_LivingAlone <- PoorFairHealth_t %>% filter(DEM_MARSTA %in% c(2,3,4)) %>% nrow()
GoodHealth_WithFamily <- GoodHealth_t %>% filter(DEM_MARSTA == 1) %>% nrow()
GoodHealth_LivingAlone <- GoodHealth_t %>% filter(DEM_MARSTA %in% c(2,3,4)) %>% nrow()
```
~~~

~~~
```{r}
# 2*2 x-tab
table = rbind(c(PoorFairHealth_WithFamily, PoorFairHealth_LivingAlone), c(GoodHealth_WithFamily,GoodHealth_LivingAlone))
rownames(table) = c("Poor Fair Health", "Good Health")
colnames(table) = c("With Family", "Living Alone")
table
```

```{r}
fisher.test(table)
```

### Q7) Loneliness and risk of depression

#### Method 1: Measure depression vs not depression as: HLT_OCDEPRSS == 1 vs. HLT_OCDEPRSS == 2
```{r}
# filter data
Depression_t <- MCBS %>% filter(ADM_H_MEDSTA == 1) %>% filter(HLT_OCDEPRSS == 1)
NotDepression_t <- MCBS %>% filter(ADM_H_MEDSTA == 1) %>% filter(HLT_OCDEPRSS == 2)

Depression_WithFamily <- Depression_t %>% filter(DEM_MARSTA == 1) %>% nrow()
Depression_Alone <- Depression_t %>% filter(DEM_MARSTA %in% c(2,3,4)) %>% nrow()

NotDepression_WithFamily <- NotDepression_t %>% filter(DEM_MARSTA == 1) %>% nrow()
NotDepression_Alone <- NotDepression_t %>% filter(DEM_MARSTA %in% c(2,3,4)) %>% nrow()
```

```{r}
# 2*2 x-tab
table2 = rbind(c(Depression_WithFamily, Depression_Alone), c(NotDepression_WithFamily, NotDepression_Alone))
colnames(table2) = c("With Family", "Living Alone")
rownames(table2) = c("Depression", "Not Depression")
table2
```

```{r}
fisher.test(table2)
```

#### Method 2: Measure depression vs not depression as: HLT_OCDEPRSS == 1 vs. HLT_OCDEPRSS !=1
```{r}
# filter data
Depression_t <- MCBS %>% filter(ADM_H_MEDSTA == 1) %>% filter(HLT_OCDEPRSS == 1)
NotDepression_t_2 <- MCBS %>% filter(ADM_H_MEDSTA == 1) %>% filter(HLT_OCDEPRSS != 1)

Depression_WithFamily <- Depression_t %>% filter(DEM_MARSTA == 1) %>% nrow()
Depression_Alone <- Depression_t %>% filter(DEM_MARSTA %in% c(2,3,4)) %>% nrow()

NotDepression_WithFamily_2 <- NotDepression_t_2 %>% filter(DEM_MARSTA == 1) %>% nrow()
NotDepression_Alone_2 <- NotDepression_t_2 %>% filter(DEM_MARSTA %in% c(2,3,4)) %>% nrow()
```

```{r}
# 2*2 x-tab
table3 = rbind(c(Depression_WithFamily, Depression_Alone), c(NotDepression_WithFamily_2, NotDepression_Alone_2))
colnames(table3) = c("With Family", "Living Alone")
rownames(table3) = c("Depression", "Not Depression")
table3
```

```{r}
fisher.test(table3)
```
~~~