Brandeis
UNIVERSITY

Advanced Topics in Healthcare Data Analytics and Data Mining

Case Study: Analytics of Patients and Consumers Survey

Erin Cao

Bo Chen

Qingyue Su

Peihan Tian

Jiujun Zhang

Kaihang Zhao

04/01/2020

Brandeis
UNIVERSITY

Introduction

Data analytics on evaluation of healthcare systems has become independent when figuring out more patient-centered and health financing suggestions. One source of data and information is the healthcare utilization survey. In this case study, we will use the 2016 US Medicare Current Beneficiary Survey (MCBS) Public Use File (PUF) to analyze various healthcare performance and relationships. Our topics will cover: racial disparity in ability to pay for care, gender differential in healthcare utilization, whether education backgrounds promote health lifestyle, a causal relationship between obesity/depression, and gender, and the health risk of loneliness.

I. Racial disparity in ability to pay for care

In theory, the responders in the MCBS are all insured by Medicare. Most of them are eligible by age meaning they are seniors who have reached the retirement age 65 which is the eligibility criterion for Medicare. In this section, we want to investigate if there is any racial disparity in terms of affordability of the out of pocket healthcare costs. We are focusing on the group of people who are eligible for Medicare only because of age >65 and two racial groups Non-Hispanic White and Non-Hispanic Black. To indicate the affordability of the pocket healthcare costs, we use the variable “ACC_HCDELAY”: Last year ever delay in care due to cost as a proxy variable. We then derive a table well illustrating the proportion of two racial groups who have delayed payment.

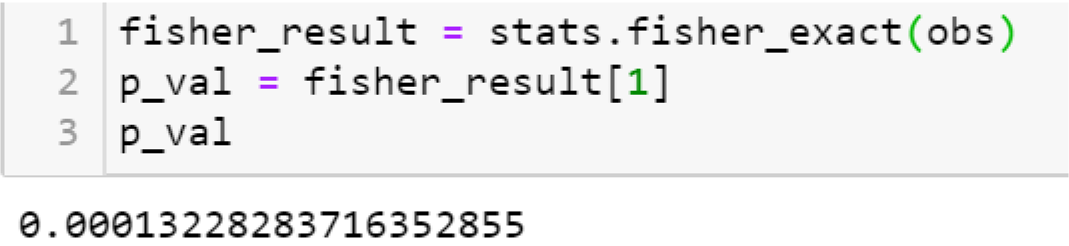
Count	Non-Hispanic White	Non-Hispanic Black
Delay (Yes)	462 (5.63%)	77 (9.07%)
Delay (No)	7737 (94.37%)	772 (90.93%)
Total	8199 (100%)	849 (100%)

Table I-1

We then performed a Fisher’s Exact test:

Null hypothesis: Race has no effect on whether a person delays payments.

Alternative hypothesis: There is a racial disparity with regard to financial difficulties.



```
fisher_result = stats.fisher_exact(obs)
p_val = fisher_result[1]
p_val
```

```
0.00013228283716352855
```

Results from Fisher’s Exact test indicate that in fact there is a highly statistically significant racial difference in terms of financial difficulties, and they are not independent of one another (p<=0.001). 9% of Non-Hispanic Black have delayed payments, while only about 5.6% Non-Hispanic White delayed their payments, which means that Black Americans are more likely than White Americans (non-Hispanic) to delay a healthcare payment.

II. Gender differentials in healthcare utilization

It is common sense that during reproductive age, women seek and need more health services than males. Therefore, in the second part, to explore the gender differentials in healthcare use fairly, we filtered individuals who are over 65 years old. We assumed the number of physician visits as a measure to healthcare utilization, and replaced the scale numbers with the midpoint of each range.

ADM_H_PHYEVT	Total office visits in current year (FFS)	Midpoint adjustment
0	No office visit	0
1	1 to 5 office visits	3
2	6 to 10 office visits	8

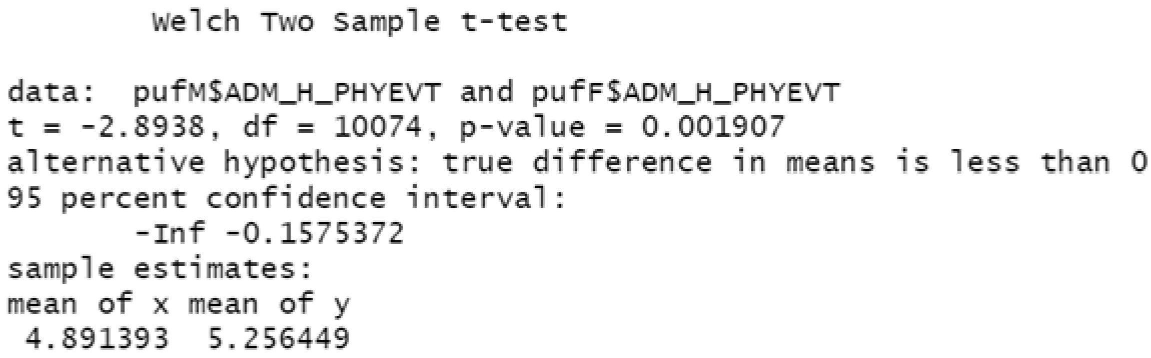
3	11 to 15 office visits	13
4	16 to 20 office visits	18
5	21 or more office visits	23

Table II-1

According to our work, we found that the number of physician visits was 4.89 visits for a male, and 5.26 visits for a female on average. To statistically examine the gender differential in healthcare utilization, a t-test was executed. The results revealed that p-value was extremely small that we can reject the null hypothesis. There was a significant difference between males and females on their physician visits and healthcare utilization. Excluding the reproduction period, the pattern that women utilize more healthcare services than men continues onto older periods (over 65 years old). It is reasonable to believe that men are relatively lazier or more ignorant than women on health seeking behavior.

Table II-2: # of Physician Visit (range midpoint)

Sex	0	3	8	13	18	23	Weighted Average
Male	2116	962	746	419	216	200	4.89
Female	2508	1258	1044	604	277	279	5.26



```
        Welch Two Sample t-test

data:  pufM$ADM_H_PHYEVT and pufF$ADM_H_PHYEVT
t = -2.8938, df = 10074, p-value = 0.001907
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
       -Inf -0.1575372
sample estimates:
mean of x mean of y 
 4.891393  5.256449 
```

III. Relationship between education and health

In dealing with the data, we consider the variable HLT_BMI_CAT values 4 and 5 as obese, and for the other values include N/A, we assume that they are all defined as healthy. By creating two categories for education and already two categories for obesity we do our Fisher's Exact test:

Null hypothesis: There is no significant relationship between education and health.

Alternative hypothesis: Education has an effect on health.

To test our hypothesis, we use data analysis to make contingency table and Fisher's Exact test, the result shows below:

Count	Obese	Healthy	Total
Low_edu	2224 (31.24%)	4895 (68.76%)	7119 (100%)
High_edu	1591 (27.75%)	4142 (72.25%)	5733 (100%)

Table III-1

P-value= 1.7639176852457522e-05<0.05, so we reject the null hypothesis.

From our results, we can conclude that education will affect health, to make a further analysis, we can see that the obese rate of highly educated people is 31.24%, while low educated is 27.75%, which means higher levels of education helps individuals to maintain their health and stay below the risky BMI levels. Folland's book theory suggested that education makes individuals healthier through better lifestyle and more informed life and health related decisions, our result shows that Folland's book theory is statistically valid.

The result can be explained as highly educated people always can get better jobs, more money and many other benefits, such as better health insurance, which leads to better access to quality health care. They would also chase for a healthy lifestyle, such as paying more on healthier diets and working out frequently.

In summary, by analyzing the data between education and health, we can see that people with high education can get a better health condition.

IV. Dependency between obesity and depression

There are studies that claims the increased risk of depression and anxiety disorder among obese individuals. To study the relationship between obesity and depression, firstly, according to HLT_BMI_CAT we divide the data into non-obesity and obesity, corresponding to 69.4% and 30.6% of our dataset. Then we count the number of depression and non-depression by HLT_OCDEPRSS. We notice that there are some unspecified values which are not 1 or 2 and we choose to drop these values.

Variable Name	Category	Value and Description
HLT_BMI_CAT	Obese Condition	1: Underweight, <18.5 2: Healthy, 18.5-<25 3: Overweight, 25-<30 4: Obese, 30-<40 5: Extreme or high risk obesity
HLT_OCDEPRSS	Depression Condition	1: Yes 2: No

Table IV-1

We make a matrix to conduct Fisher Exact test to see if the difference in the depression rate is in fact significantly different in the two groups.

Null hypothesis: The relative proportions of obesity are independent of depression.

Alternative hypothesis: The two variables are dependent on each other.

	Depression	Non Depression	Total
Obesity	1366 (11.0%)	2446 (19.7%)	3812 (30.6%)
Non Obesity	2039 (16.4%)	6593 (53.0%)	8632 (69.4%)
Total	3405 (27.4%)	9039 (72.6%)	12444 (100%)

Table IV-II

Fisher Exact test results in a p-value of 8.52e-44 (<0.05) which is extremely small. In this case, we reject the null at 5% confidence level. We would say that there is a relationship between obesity and depression. Healthy people tend not to have depressive disorder.

V. Are obesity and depression related?

The previous research above revealed that there is a positive relationship between obesity and depression, without getting into complexities involving statistical examination of reciprocal causation. However, after reading some literature essays investigating the two sides of the obesity-depression causal relationship, we come up with a guess that gender may be one of the causes in this reciprocal causation. First, take a look at these three category variables.

Variable Name	Category	Value and Description
DEM_SEX	Gender	1: male; 2: female

HLT_BMI_CAT	Obese Condition	.: Inapplicable/Missing; 1: Underweight, <18.5;
		2: Healthy, 18.5-<25; 3: Overweight, 25-<30;
		4: Obese, 30-<40; 5: Extreme or high risk obesity

HLT_OCDEPRSS	Depression Condition	D: Don't know; R: Refused; .: Inapplicable/Missing; 1: Yes; 2: No

Table V-1 Description of variables

The description table above shows the specific value of these three category variables and the corresponding meaning. To make our analysis clear and easy to understand, we make two specific assumption according to obese condition and depression condition here：

- About obese condition, we assume that people with obese and extreme or high-risk obesity (HLT_BMI_CAT = 4 and 5) represents all the obese population, while people without obesity (HLT_BMI_CAT = 1, 2 and 3) represents healthy people by ignoring the missing value.
- About depression condition, we assume that people with depression (HLT_OCDEPRSS = 1) represents all the depressed population, while people without depression (HLT_OCDEPRSS = 2) represents undepressed people by ignoring the missing, do not know and the refused value.

Therefore, based on our assumption, we started to dig into the relationship between gender and obese condition, and that between gender and depression condition separately at first. Below are some outputs.

(1) Gender & Obese Condition

After analyzing the relevant data in R, we found that males have a slightly less risk of being obese (7.91%) than females. Although this difference is not that huge, from the result of Fisher's exact test, we can still tell that the alternative hypothesis that there is a correlation between gender and obesity should be accepted.

Gender	Obese Population	Healthy Population	Total
Male (% within gender)	1682 (39.49%)	4022 (70.51%)	5704 (100%)
Female (% within gender)	2133 (31.58%)	4622 (68.42%)	6755 (100%)

* Rate differential in male: 31.02%

* Rate differential in female: 36.84%

TableV-2 Gender - Obese Condition Cross Table

	Obese Condition	
Gender	Odds Ratio	P value
	0.906	0.012

TableV-3 Fisher test on Gender - Obese Condition

(2) Gender & Depression Condition

The outcome revealed that the male has significantly lower risk of suffering depression (9.46%) than the female, which also makes sense. Since women have to overcome many difficulties in their lives, such as childbearing, child rearing, and so on. Thus, it is reasonable for women to be more likely to suffer depression when compared with men. It is also verified considering the outcome from Fisher's exact test.

Gender	Depressed Population	Undepressed Population	Total
Male (% within gender)	1293 (22.34%)	4496 (77.66%)	5789 (100%)
Female (% within gender)	2238 (31.80%)	4799 (68.20%)	7037 (100%)

* Rate differential in male: 55.32%

* Rate differential in female: 36.40%

TableV-4 Gender - Depression Condition Cross Table

	Depression Condition	
Gender	Odds Ratio	P value
	0.617	3.57 * e-33

TableV-5 Fisher test on Gender - Depression Condition

After clarifying that gender has some effect, we want to testify this outcome in another way. Therefore, we repeated the same procedures in the previous section twice separately for male and female subjects and see which gender has a higher rate differential.

(1) Male: Obese Condition & Depression Condition

When we test the relationship between depression condition and obese condition in the male sub-set, we figured out that the depressed population has significantly higher risk of suffering depression (11.79%) than the undepressed, which is the same as what we got in the previous section. It is also testified based on the outcome from Fisher’s exact test.

Depression Condition	Obese Population	Healthy Population	Total
Depressed Population (% within Depression Condition)	491 (38.66%)	779 (61.34%)	1270
Undepressed Population (% within Depression Condition)	1189 (26.87%)	3236 (73.13%)	4425

* Rate differential in depressed population: 22.68%

* Rate differential in undepressed population: 46.26%

TableV-6 Obese Condition - Depression Condition Cross Table (Male)

	Obese Condition	
Depression Condition	Odds Ratio	P value
	1.715	1.44 * e-15

TableV-7 Fisher test on Depression - Obese Condition (Male)

(2) Female: Obese Condition & Depression Condition

After analyzing the relationship between depression condition and obese condition in the male sub-set, we repeated that procedure in the female sub-set. It was discovered that the depressed population has significantly higher risk of suffering depression (13.74%) than the undepressed, which is the same as what we got in the previous section and is larger than what we got in the male population. This is also testified based on the outcome from Fisher’s exact test. Therefore, we can reasonably infer that gender is a determinant to the relationship between the reciprocal causation of obese condition and depression condition.

Depression Condition	Obese Population	Healthy Population	Total
Depressed Population (% within Depression Condition)	875 (40.98%)	1260 (59.02%)	2135
Undepressed Population (% within Depression Condition)	1257 (27.24%)	3357 (72.76%)	4614

* Rate differential in depressed population: 18.04%

* Rate differential in undepressed population: 45.52%

TableV-8 Obese Condition - Depression Condition Cross Table (Female)

	Obese Condition	
Depression Condition	Odds Ratio	P value
	1.855	6.45 * e-29

TableV-9 Fisher test on Depression - Obese Condition (Female)

VI-1. Loneliness and health

The proportion of people with poor health is greater among individuals who live alone in comparison to those who stay with family. The data we analyzed shows that there were 20.6% individuals with poor health living alone compared to 16.3% individuals with poor health among those who lived with their family. We also conducted a Fisher test to ascertain whether there is a significant difference between living and health conditions.

	WithFamily	LivingAlone	Total
PoorHealth	895(16.3%)	1049(20.6%)	1944
GoodHealth	4590(83.7%)	4033(79.3%)	8623
Total	5485	5082	10567

Table VI-1 Living-Health Conditions Cross Table

	Fisher Test
P-value	*1.138e-08*
Odds Ratio	*1.33894*

Table VI-2 Fisher test on Living - Health Conditions

As the table shows, the p value is significant at the 5% level, which means that the true odds ratio is not equal to 1 and we can reject the null hypothesis.

The odds ratio is 1.33894 and it means that the probability of people having poor health if they live alone are 33.894% higher than people living with their family members.

As a result, we can conclude that living conditions can definitely affect individuals' health conditions. Perhaps there are other issues that might affect health conditions and need further consideration.

VI-2. Loneliness and risk of depression

The table below shows that depression is more prevalent among individuals who stay alone when compared to those who stay with their family members. The data we analyzed shows that there were 24.6% individuals having a depression compared to 18% of those staying with their family members. This analysis aimed at determining whether this observed difference was random or due to random choice. To solve this concern, we did a Fisher Test.

	WithFamily	*LivingAlone*	*Total*
Depression	*992 （18%）*	*1255 （24.6%）*	*2247*
NoDepression	*4518 （82%）*	*3854 （75.4%）*	*8372*
Total	*5510*	*5109*	*10619*

Table VI-3 Living-Depression Conditions Cross Table

	Fisher Test
P-value	*2.2e-16*
Odds Ratio	*1.483*

Table VI-4 Fisher test on Living - Depression Conditions

As the table shows, the p value is significant at the 5% level, which means that the true odds ratio is not equal to 1 and we can reject the null hypothesis.

The odds ratio is 1.487 and it means that the probability of people having a depression if they live alone are 48.3% higher than people living with their family members.

In conclusion, we can say that people living with family are more likely to be healthy and have a happy mood than those without a close family member.