**Factors that Contribute to Healthcare Cost**

Hao Ni, Dading Shi, Hyesun Jun, Sein Lee

Emory College of Arts & Sciences

QTM302W: Technical Writing

Dr. David Fisher

Apr. 16, 2020

**Abstract**

Skyrocketing costs in healthcare have been a headache for many people in the United States for a long time. People are concerned about their safety, but also daunted by the ridiculous cost. Our study seeks to determine which factors contribute most to yearly healthcare cost spent per year using a Kaggle dataset that documents 1,338 individuals’ yearly healthcare costs spent per year and their demographic information including their gender, smoking status, age, region, number of children, and BMI index in 2017 in the United States. The dataset has allowed this study to investigate the most influential factor behind healthcare costs, which we hypothesized would be smoking. The Previous research has shown that healthcare cost for smokers is as much as 40% higher than those for non-smokers depending on age. However, this outdated research had only compared smokers to non-smokers, so we aimed to take other variables into account. Before conducting the multiple linear regression, we checked a set of assumptions including multicollinearity and heteroscedasticity. Out of the variables, smoking had the highest t-value of 57.72 in comparison to age and BMI. This agrees with our hypothesis that smoking is the biggest contributing factor to personal healthcare costs. Our findings may imply that smoking is the cause of many health problems that require medical attention, thus emphasizing the need to abstain from smoking in order to decrease the risk of illness, therefore reducing healthcare costs. It may also be a possibility for the FDA to regulate nicotine levels in cigarettes in order to reduce harm for the public.

*Keywords*: healthcare cost, smoking, insurance cost, medical insurance cost, public health

**Introduction**

Skyrocketing costs in healthcare, defined by an individual’s expense on medical services over a year, have been a headache for many people in the United States for a long time. Over the recent few months, the impact of these costs has become even more apparent due to the quick-spreading coronavirus. According to healthline.com, the cost of an initial test for coronavirus is $3,270. This price is often too high for people without health insurance to afford. People are concerned about their safety, but also daunted by the ridiculous cost.

Nowadays, the medical costs in the United States ranks higher than any other country in the world, with 3 trillion dollars of spending each year. In the 1960s, the share of gross domestic product (GDP) accounted for by health care was only 5 percent of the U.S. economy. Since the late 1990s, health care spending has increased at a faster rate that of the GDP, population, and inflation. By 2018, the spending has increased to 17.7 percent of GDP, which was three times more than that of the 1960s (“Effects of Spending,”, 2017). It is reported that on average, a U.S. household suffers from a huge financial burden of $8,000 annually because of lost wages, higher premiums, taxes, and additional out-of-pocket expenses (Epstein, 2019).

Despite the rising healthcare costs, there is some hope for assistance through a health insurance plan, which significantly reduces the burden of non-insured medical expenses. Thus, robust health insurance becomes a necessity for every household to avoid surprisingly high costs of medical emergencies.

Personal expenditure on health services varies among different people with different backgrounds, and variation invites scrutiny about how, exactly, these costs are determined. Previous research that explored this topic. (Tsai et al. (2011) found that the 2008 per‐person direct medical cost for those who were overweight was $266 and $1,723 for those who were obese. The aggregate national cost of being overweight and obese was $113.9 billion. Depending on the source of total national healthcare expenditures used, the direct medical cost of overweightness and obesity combined is approximately 5.0% to 10% of US healthcare spending. Smoking is also a crucial factor in health insurance rate. The health care cost for smokers is as much as 40% higher than those for non-smokers depending on age. In 1992, the Surgeon General reported that the estimated average lifetime medical costs for a smoker exceed those for a nonsmoker by more than $6,000 (Barendregt et al., 1997). However, this research had only compared smokers to nonsmokers, so we plan to take other variables into account.

  Our study seeks to determine which factors contribute most significantly to the yearly healthcare of cost spent by an individuals per year using a Kaggle dataset that documents 1,338 individuals’ medical insurance costs and their demographic information including their gender, smoking status, age, region, number of children, and BMI index in 2017 in the United States. The dataset has allowed this study to investigate the most influential factor behind healthcare cost, which we hypothesized the factor contributing most to this cost would be smoking.

**Methods**

**Data**

In order to determine which factors contribute most to the yearly healthcare cost for individuals using a Kaggle dataset that documents 1,338 individuals’ healthcare costs and their demographic information including their gender, smoking status, age, region, number of children, and BMI index in 2017 in the United States. This dataset had no missing values, so we did not need to apply any data reduction process or missing value interpolation. Moreover, all 1,338 individuals were randomly selected from the population pool, thus causing no selection bias.

**Variables**

**Age**. We transformed this variable by binning ranges of age. We classified ages 18 to 24 as 1, ages 25 to 34 as 2, ages 35 to 44 as 3, ages 45 to 54 as 4, ages 55 to 64 as 5, and ages 65 or older as 6. The participants' ages were fairly evenly distributed throughout the age groups. Provide percentages to make your descriptive statistics complete.

**Sex.** This variable shows whether the insurance contractor is female or male. We classified the dichotomous results as 1 for male and 0 for female. Out of all the participants, 50.5% (676) were male and 49.7% (662) were female.

**BMI.** BMI conveys whether the weights are relatively high or low relative to height. The objective index of body weight (kg / m ^ 2) uses the ratio of height to weight, ideally 18.5 to 24.9 according to the American Cancer Society.

**Children.** The number of children covered by health insurance / Number of dependents is another variable to see how this variable affects the medical cost. Within the 1,338 participants, 67% (898) of the individuals had 0 to 1 dependents, 30% (397) had 2 to 3, and 3% (43) had 4 to 5 dependents.

**Smoker.** The major variable that we have focused on was whether the participant smokes or not. We hypothesized that smoking would be the biggest contributor to healthcare costs. We classified the results dichotomously as 1 for smokers and 0 for non-smokers. Within the 1338 individuals, around 79.5% (1064) of the participants were nonsmokers.

**Region.** This variable indicates the beneficiary's residential area in the US: NorthEast, SouthEast, SouthWest, NorthWest . Within the total number of individuals, 24% (324), 27% (364), 24% (325), 24% (325) were where the participants lived respectively.

**Charges.** This variable represents medical costs paid per year by an individual , which is the outcome variable. Within the participants, 53% (712) paid less than $10,000 a year for medical expenses.

**Table 1**

*Demographics of participants including smoking status, age, sex, obesity status (BMI), number of children/dependents, region, and charges.*

|  |  |  |
| --- | --- | --- |
| **Category** | **Level** | **n, %** |
|  | All | 1338, 100% |
| **Smoking status** | Nonsmoker | 1064, 79.5% |
|  | Smoker | 274, 20.5% |
| **Age** | Age 18 to 24 | 278, 20.7% |
|  | Age 25 to 34 | 271, 20.2% |
|  | Age 35 to 44 | 260, 19.4% |
|  | Age 45 to 54 | 287, 21.4% |
|  | Age 55 to 64 | 242, 18.1% |
| **Sex** | Male | 676, 50.5% |
|  | Female | 662, 49.7% |
| **Obesity status** | Underweight | 41, 3% |
|  | Normal | 204, 15% |
|  | Overweight | 386, 29% |
|  | Obese | 707, 53% |
| **Children** | 0-1 | 898, 67% |
|  | 2-3 | 397, 30% |
|  | 4-5 | 43, 3% |
| **Region** | Southeast | 364, 27% |
|  | Southwest | 325, 24% |
|  | Northeast | 324, 24% |
|  | Northwest | 325, 24% |
| **Charges** | $0-10,000 | 712, 53% |
|  | $10,000-20,000 | 353, 26% |
|  | $20,000-30,000 | 111, 8.3% |
|  | $30,000-40,000 | 83, 6% |
|  | $40,000-50,000 | 72, 5% |
|  | $50,000-60,000 | 4, 0.3% |
|  | $60,000+ | 3, 0.2% |
|  |  |  |

**Methods**

        The primary hypothesis, that smoking is the biggest contributor to healthcare costs, will be tested through by fitting a multiple linear regression model. Before using a multiple linear regression, We checked and followed the following steps before using a multiple linear regression we ensured our data met the assumptions of this method:.

1. The relationship between independent and dependent variables were linear (MLR 1)
2. The subjects were selected randomly (MLR 2)
3. There were no perfect collinearity between the explanatory variables (MLR 3)
4. The error had zero expected value for independent variables (MLR 4)
5. Personal health cost had the same finite variance across all factors (MLR 5).

If these factors were heteroskedastic, which we predicted they are, then we used a linear regression model with robust standard errors. All assumptions were satisfied as we checked. Therefore, we can perform regression analysis with our data.

**Results**

To establish the relationship between smoking and personal healthcare costs, we ran multiple linear regressions. The variables we analyzed included age, BMI, number of children, which were numerical variables, and sex, smoker, and region, which were categorical. The criterion variable of this analysis was the healthcare cost, which was how much a person pays for healthcare in a given year. Before conducting the multiple linear regression, we checked a set of assumptions including multicollinearity and heteroscedasticity. Since we had multiple categorical variables, we had to calculate the variance inflation factor directly in order to assess multicollinearity. No variables exceeded the VIF value of 10, thus meeting the assumption that all of the variables are not significantly correlated to each other.

The factors that contributed the most to an increase in healthcare cost were age, BMI, whether or not one had children, and smoking. Out of these variables, smoking had the highest t-value of 57.72 in comparison to age which had 21.59 and bmi which had 11.86. This agrees with our hypothesis that smoking is the biggest contributing factor to personal healthcare costs. \*Results showed that the linear combination of predictor variables accounted for 74.9% of the variance in healthcare costs (Adjusted R2 = .749, F(8,1329) = 500.81, p < .001) Amongst the eight variables, age, BMI, and smoker had significant relations to healthcare costs (Table 2).

**Table 2**

*Results of the Regression Model: Factors of Healthcare costs*

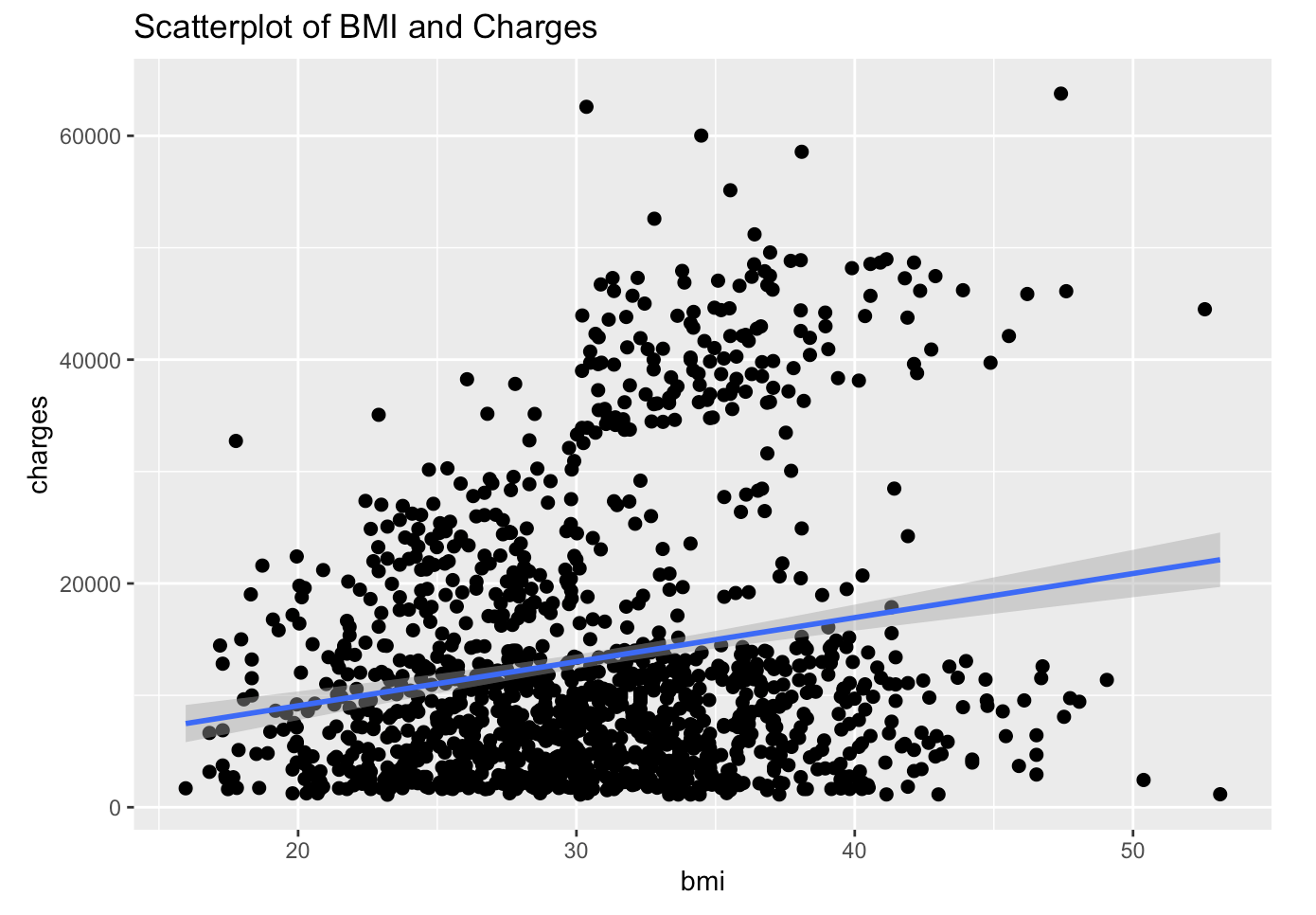
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Predictor Variable       Estimate   Std .Error          t-value             p-value** | | | | |
| (Intercept) | -19938.5 | 987.8 | -12.086 | < .001 \*\*\* |
| Age | 256.9 | 11.9 | 21.587 | < .001 \*\*\* |
| Sex (male) | -131.9 | 332.9 | -0.394 | 0.693 |
| BMI | 339.2 | 28.6 | 11.860 | < .001 \*\*\* |
| Children | 475.5 | 137.8 | 3.451 | 0.006 |
| Smoker (yes) | 23848.5 | 413.1 | 57.723 | < .001 \*\*\* |
| Region (Northwest) | -353.0 | 476.3 | -0.741 | 0.459 |
| Region (Southeast) | -1035.0 | 478.7 | -2.162 | 0.031 \* |
| Region (Southwest) | -960.0 | 477.9 | -2.009 | 0.048 \* |
| Note N = 1338, Model R2 = 0.7509, F(8,1329) = 500.81, p < .001, Adjusted R2 = 0.7494, \*p < .05, \*\*p < .01, \*\*\*p < .001 | | | | |

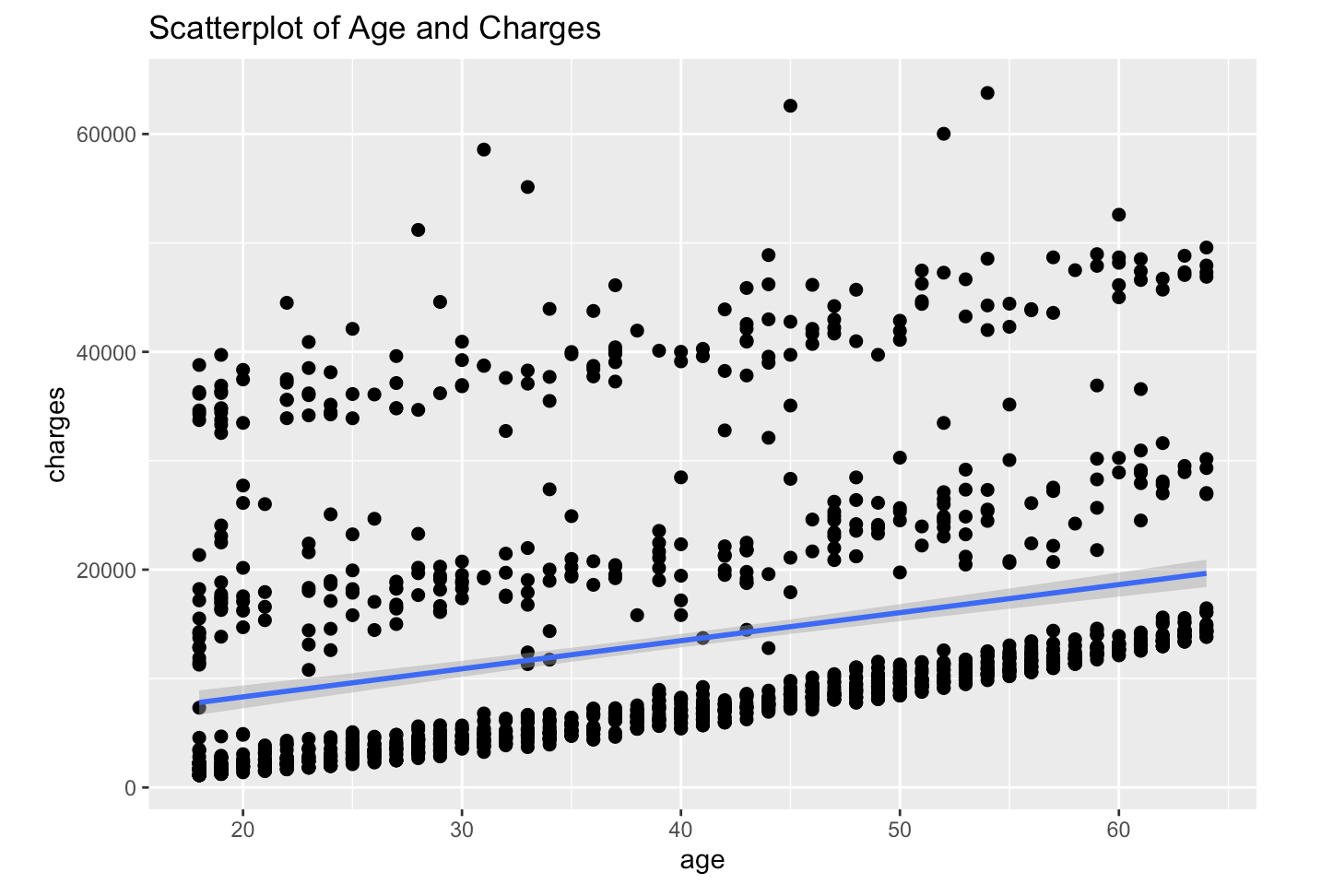
 Among all six regressors, only Age, BMI, and Smoker show a significant relationship to the model since all of them have a p-value smaller than 0.001, and we are going to deliver an interpretation of these three variables. Holding all other variables constant, a smoker will pay $23,848.50 more than nonsmokers in medical cost on average. Similarly, an increase of one year old in age will lead to a $256.90 increase in medical cost on average. Last but not least, an increase in one point of BMI will bring up the medical cost by $339.20 on average.

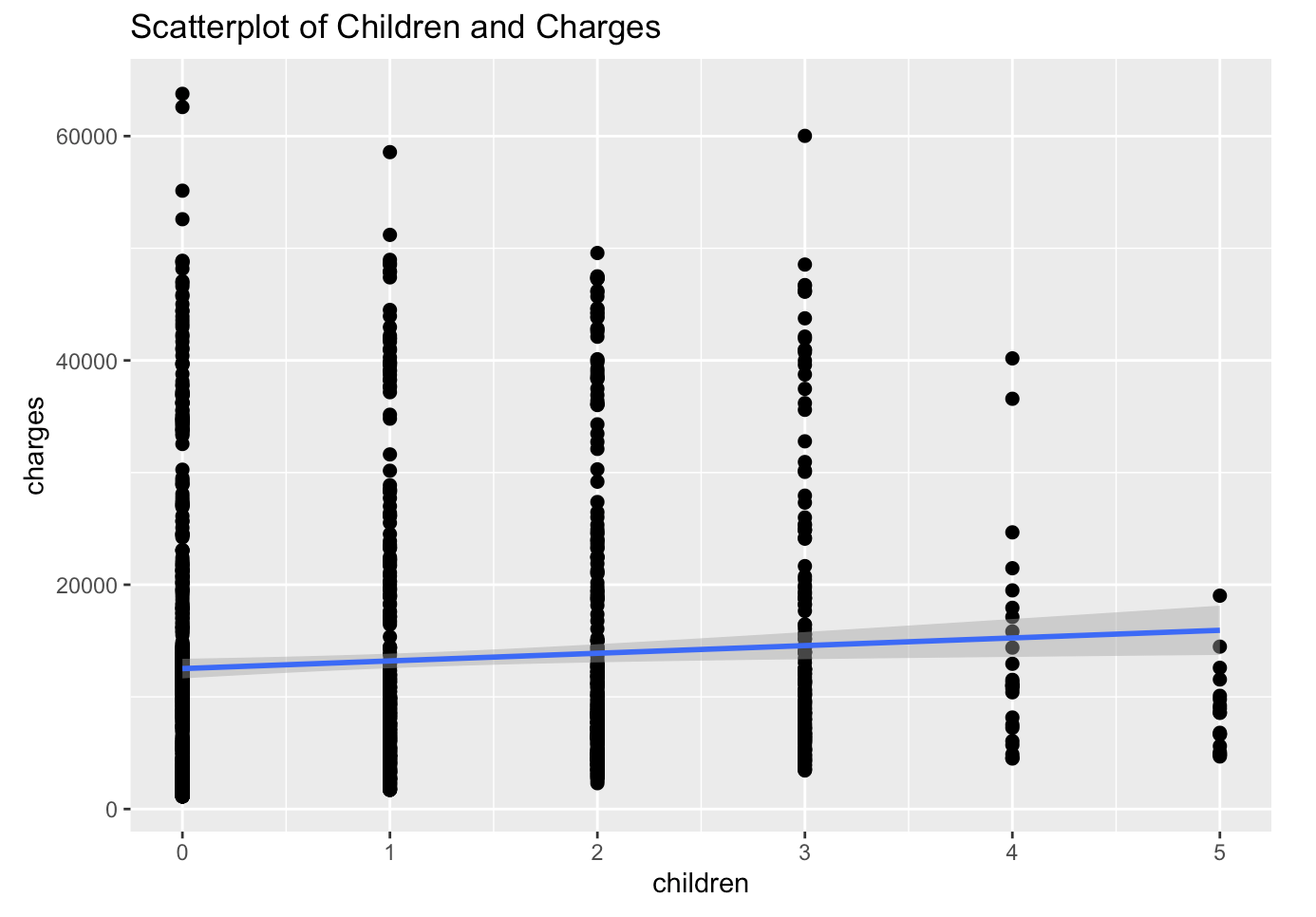
         The scatter plots and histogram were constructed to evaluate the relationship between the healthcare costs and the variables: age, BMI, children (Figure 1). To be specific, the scatter plots were constructed to test the linearity between the variables and the charges.

**Figure 1**

*Scatterplot of BMI and Charges, Age and Charges, Children and Charges*







**Discussion**

         Even though factors such as old age or a high BMI are significantly associated with increases in personal healthcare spending, our research identifies that smoking is the biggest contributor. This is consistent with previous research that the health care cost for smokers is as much as 40% higher than those for nonsmokers depending on the age (Barendregt et al., 1997). However, this outdated research had only compared smokers to non-smokers, so we aimed to take other variables into account.

Our findings may imply that smoking is the cause of many health problems that require medical attention, thus emphasizing the need to abstain from smoking in order to decrease the risk of getting sick, therefore reducing healthcare costs. This makes it crucial for adolescents to receive education about the detriments of smoking before entering adulthood, as well as a reliable support program for adults who are trying to quit smoking. It may also be a possibility for the FDA to regulate nicotine levels in cigarettes in order to reduce harm for the public.

A possibility for further research is comparing amongst smokers which health problem related to smoking leads to more healthcare costs, and even comparing the smoking habits of the smokers to see whether the frequency or amount of smoking each week shows a difference in personal healthcare spending. It would also be interesting to investigate the healthcare costs for people who are non-smokers but have been exposed to secondhand smoke and relate it to previously confirmed health problems caused by secondhand smoke.

**References**

Barendregt, J. J., Bonneux, L., & van der Maas, P. J. (1997). The Health Care Costs of Smoking.

*New England Journal of Medicine, 337(15)*, 1052–1057.

<https://doi.org/10.1056/NEJM199710093371506>

Effects of Health Care Spending on the U.S. Economy (2017, February 21). *ASPE*

<https://aspe.hhs.gov/basic-report/effects-health-care-spending-us-economy#_edn12>

Epstein, L. (2020, January 29). 6 Reasons Healthcare Is So Expensive in the U.S. Retrieved from

https://www.investopedia.com/articles/personal-finance/080615/6-reasons-healthcare-so-expensive-us.asp

Tsai, A. G., Williamson, D. F., & Glick, H. A. (2011). Direct medical cost of overweight and

obesity in the USA: A quantitative systematic review. *Obesity Reviews, 12(1)*, 50–61.

<https://doi.org/10.1111/j.1467-789X.2009.00708.x>