Common Risk Factors and their Correlation to Diabetes in Americans

DATA363 Final Report

Orhan Emir Gozutok and Jesus Arias

Dec 12, 2022

Introduction

Diabetes is one of the most prevalent global diseases, impacting more than 450 million people worldwide and in America. Diabetes is a serious condition in which blood sugar regulation is affected in the pancreas by hormones like insulin. Diabetes can be deadly since it is important for organs in the body to maintain a stable intake of nutrients, which largely is mediated by glucose. Without consistent presence of glucose, there have been reports of brain or other types of tissue damage. The dataset can be accessed here: Diabetes Health Indicators (link).

As new data gets published and as years progress, processing becomes better, it is important to reassess what is known in the field of medicine about diabetes and test it statistically to ensure that our knowledge across different datasets and methods produce the same results about diabetes which affects millions everyday. Also, this BRFSS survey is weighted differently than the previous ones, hence the importance of our study.

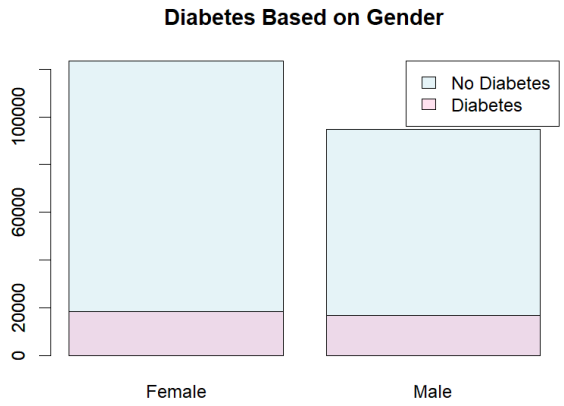
Methods

This dataset contains the 2015 annual survey results carried out by the CDC which is titled Behavioral Risk Factor Surveillance System. This program collects data from over 400,000 Americans on health metrics, on whether they have chronic conditions, preventative measures used against them, and if they have diabetes.

This data was collected from adults with a cellular phone and is residing in a private residence or college housing. A point of bias could be a selection bias against people who would not want to take the survey over the phone if offered, also because of its length due to many variables involved. The CDC states that the majority of households use some sort of phone line, whether it be a landline or a cell phone. Nevertheless, this potential source of bias is unlikely to be significant because we are looking at what affects diabetes. Our study might have been affected if we were looking at prevalence of diabetes in different age groups, where younger people would take the phone survey less.

The CDC does not report any other potential biases about this survey.

Results



Description: Diabetic vs non-diabetic based on gender (male/female)

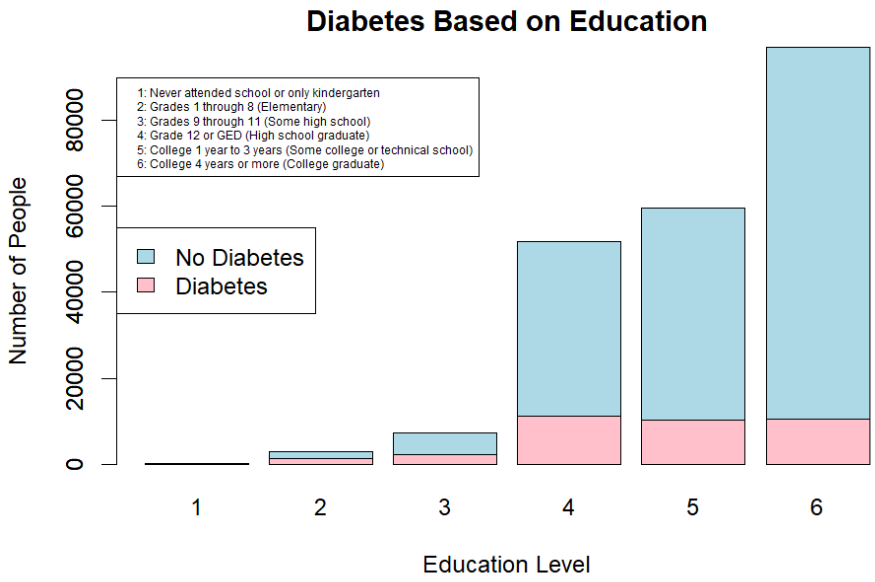
Out of everyone that took the survey, 14.9% of the females were diabetic while 17.88% of males were reported to be diabetic. As presented in the table above, there are more females than males in this study.

To confirm that this ratio of males vs females who have diabetes is not due to chance and is significant, we carry out a chi-square test, based on female vs male and diabetic vs non diabetic.

```
> genderdf
            male    female
no diabetes 123563   94771
diabetes     18411   16935
> chisq.test(genderdf)
      Pearson's Chi-squared test with Yates' continuity correction

data:  genderdf
X-squared = 250.41, df = 1, p-value < 2.2e-16
```

As seen from the chi-square independence test, diabetes is not independent of gender with a p-value of less than 2.2*10^-16, which is very significant. We see that there is a very low likelihood that we got these results based on pure chance and conclude that diabetes has a differential effect based on gender, and it can also be concluded that females are more likely to get diabetes.



Description: Diabetic vs non-Diabetic based on education level.

Of the number of people surveyed there was a clear separation between education level of graduate-plus versus some high school-below. As seen above, there are many more non-diabetic people in the survey, therefore the ratio of non-diabetic/diabetic is much more important to understand correlation between the two based on their respective education level. The ratios are shown below all as (diabetic-education) / (non-diabetic-education).

1: Never attended school or only kindergarten - 0.3700787

2: Grades 1 through 8 (Elementary) - 0.4136364

3: Grades 9 through 11 (Some high school) - 0.3196881

4: Grade 12 or GED (High school graduate) - 0.2141088

5: College 1 year to 3 years (Some college or technical school) - 0.1738532

6. College 4 years or more (College graduate) - 0.1072995

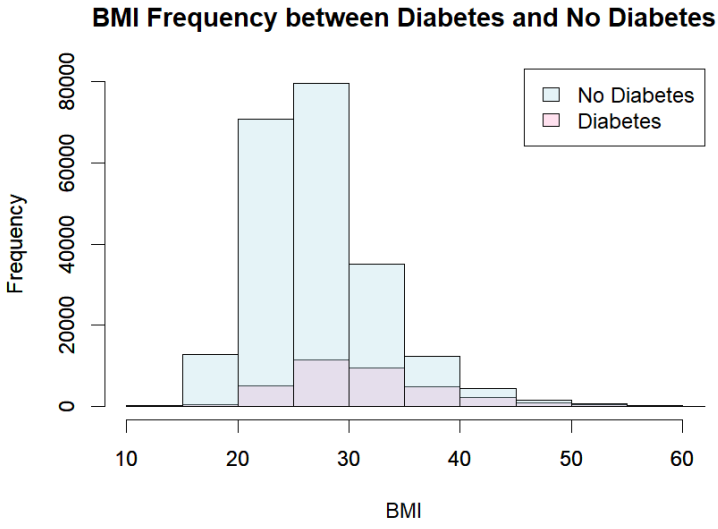
The highest ratio of diabetic/non-diabetic is of education levels 1, 2, and 3; 0.37, 0.413, 0.319 respectively. These ratios are extremely high but may be due to the extremely small population of the education levels as seen in the box plot above, and due to adults being surveyed for the study. In the education levels of 4, 5, and 6, the ratios are very similar and it may be concluded that the majority of diabetic people have a very high education level, possibly pointing to the correlation between age and education (and age vs diabetes).

```
> t.test(edYesDiabetes,edNoDiabetes)

      Welch Two Sample t-test

data:  edYesDiabetes and edNoDiabetes
t = -58.979, df = 45302, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.3660545 -0.3425073
sample estimates:
mean of x mean of y
 4.745516  5.099797
```

As seen from the results, we have two groups with significantly different means than each other. It is also an interesting but minor result that people with diabetes tend to be less educated than those with no diabetes.



Description: A histogram of BMI between people with confirmed diabetes and people with no diabetes.

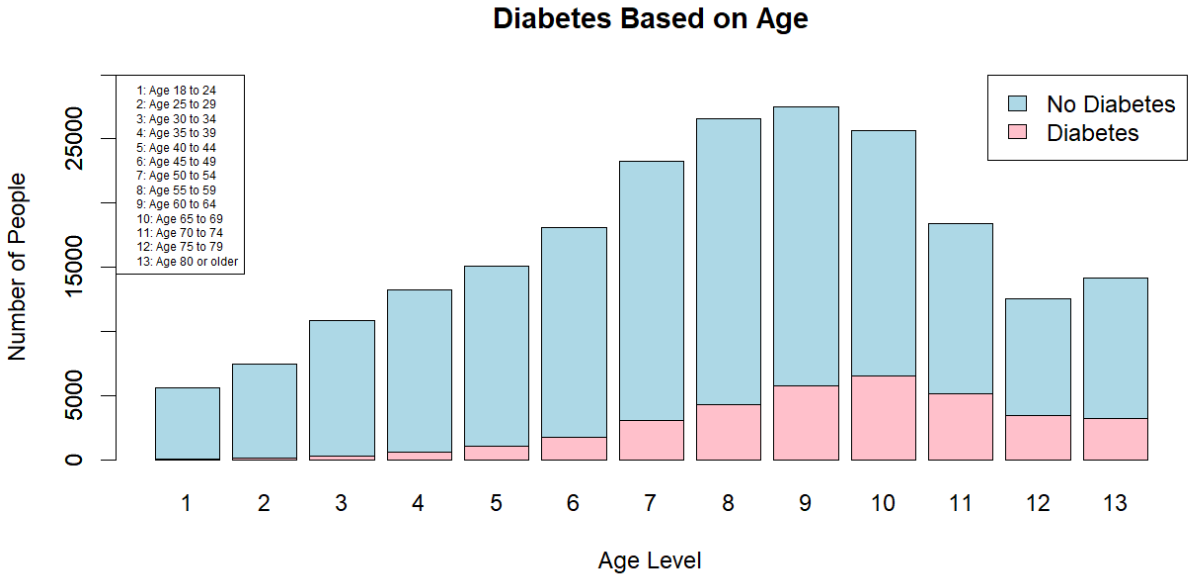
The mean BMI for non-diabetics are 27.81 while the mean BMI for diabetics are 31.91. This means that we can hypothesize that higher BMI contributes to diabetes.

To test this hypothesis and the significance of these preliminary results, we have used a t-test.

```
> t.test(bmiYesDiabetes,bmiNoDiabetes)

        Welch Two Sample t-test
data:  bmiYesDiabetes and bmiNoDiabetes
t = 99.92, df = 44093, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 4.057065 4.219416
sample estimates:
mean of x mean of y
 31.94401  27.80577
```

As seen, the t-test tells us that we have a p-value that is very small and significant that is indicating the means of BMI for diabetic and non-diabetic are different. The BMI tends to be higher for diabetic people, therefore we can say someone might be at higher risk for diabetes if they have a higher body-mass index (BMI).



Description: Bar plot of Diabetes/No Diabetes Based On Age

This bar plot highlights the extremely noticeable observation that diabetes is much more common at age groups 8, 9, 10, and 11. Therefore a person from the age of 50 - 69 has an added ratio of .9 but a person from the age of 18 - 39 has an added ratio of .1. This means that you are 9x more likely to suffer from diabetes from the ages of 50-69 than someone from the ages of 18-39.

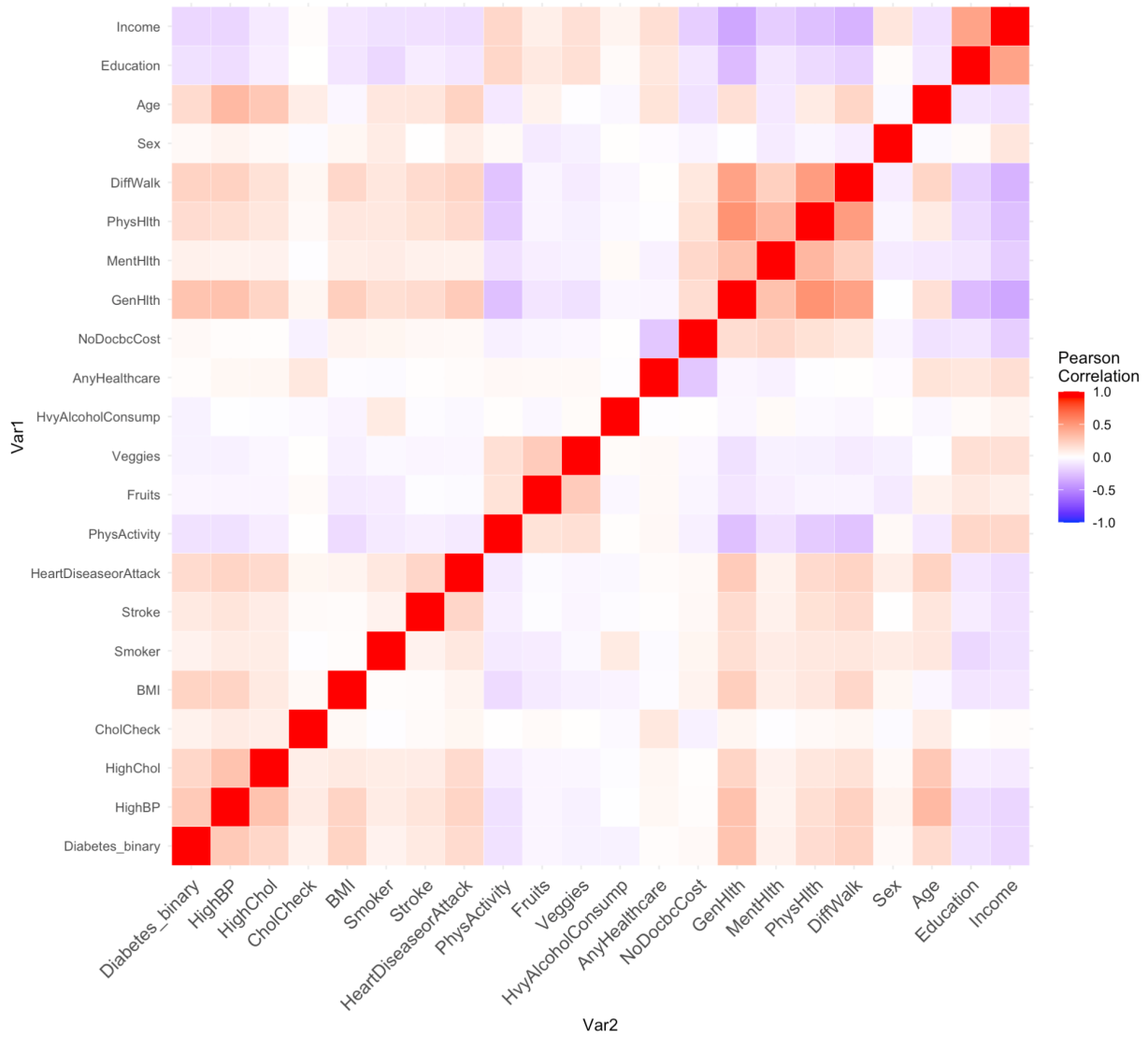
To test the significance of the difference between non-diabetic and diabetic groups' age fractions, we ran a t-test:

```
> t.test(ageYesDiabetes,ageNoDiabetes)

        Welch Two Sample t-test

data:  ageYesDiabetes and ageNoDiabetes
t = 111.31, df = 57752, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.537431 1.592546
sample estimates:
mean of x mean of y
 9.379053  7.814065
```

We can see from our t-test that diabetic and non-diabetic people have significantly different age averages, with diabetic people tending to be older. Therefore, we can say that one might be at higher risk if they are older. This is also true when we think about how age is positively related to our other variables high blood pressure and high cholesterol, which are also positively correlated with diabetes that can be seen from the heatmap below:



Description: A correlation heatmap between 22 variables in the dataset (after going through 330 variables and omitting the ones not known to contribute to diabetes, to make analysis easier).

Looking at this heatmap for the row of Diabetes_binary (0 for non-diabetic, 1 for diabetic) we can see with a quick glance that the variables that are most likely to have significant positive correlations with diabetes (shades of red, preferably darker), could be age, physical health, BMI, heart disease/attack history, hypercholesterolemia and hypertension.

Other variables negatively correlated with diabetes seem to be income, education, consuming fruits and vegetables, and physical activity.

Looking at this heatmap, we can say that there are no variables that are perfectly positively or negatively correlated with diabetes. But, we have investigated multiple variables

that when combined, might give a clearer picture about whether someone is more likely to have diabetes.

Discussion

This study has found several factors that may correlate with having diabetes. These factors are Age, BMI, Education, and Gender. These factors also have significantly different means when separating between non-diabetic and diabetic. Overall we saw different levels of plausible correlation with the simplest way being the ratio between the factor and if they have diabetes or not. However, this may not always lead to accurate results because of the importance of sample data size. It is not wise and correct to draw such conclusions with small sample sizes when separated into intervals such as education level, which contains an extremely low sample size of education levels 1-4 in the education vs diabetes bar plot. This may present some sort of participation bias that was not presented in the introduction of the data set, or is simply because adults with access to phones who participated in the survey tended to be educated by at least some amount. Overall, our results suggest that these four factors studied in this report are possible positive correlations to diabetes are presented by our heatmap displaying correlations between variables.

As mentioned previously, there was a noticeable disparity between surveys of participants with lower levels of education and age. Of course it is due to the unbiased nature of the survey to have an unknown result, however this data would be more accurate and improved upon by broadening the entry conditions of the survey. Also, what could be done to improve this study is to attempt to have the same amount of age groups, gender groups and education groups surveyed for more uniform sampling. This would allow for better analysis of how the different factors affect their vulnerability to diabetes.

A followup to this study is the addition of what state/location in the United States they are a part of. This would allow for the creation of a regional map of possible separation of diabetes based on region of residence. I suspect we would see hotspots for diabetes and dispersion from there on, as well as large areas of low rates of diabetes in specific regions based on BMI.

REFERENCES

Original Data

Centers for Disease Control and Prevention (CDC) (2015). *Behavioral Risk Factor Surveillance System Survey Data.* *https://www.cdc.gov/brfss/annual_data/annual_2015.html*

Data Set Formatting

Teboul, Alex. (2021, November). *Diabetes Health Indicators Dataset Notebook.* Retrieved September 20, 2022 from https://www.kaggle.com/code.alexteboul/diabetes-health-indicators-dataset-notebook/notebook

APPENDIX

R Version: 4.2.1

CODE

Code available on Github: https://github.com/oegozutok/diabetes

(.R file and and csv files inside project folder)