What Causes Heart Diseases in Modern Life and the Model to Predict it

A Data Analysis Project Based on CDC Health Data, 2020

Yichen Duo

2022-04-06

Contents

Abstract

This article is based on the 2020 CDC (Center of Disease Control) (LLC 2022) health survey data of 400k adults, the source of which is the Behavioral Risk Factor Surveillance System (BRFSS), as part of the health questionnaire for American residents. This paper is done in R using R code(R Core Team 2020) and produced with (Xie 2014). The purpose of this paper is to intercept the relevant causes of heart disease and related influencing factors in the survey data, to conduct data sorting, classification, and modeling analysis, to analyze the main causes of acquired heart disease, and to establish a system that can be used to predict individual Statistical models of heart disease risk.

Introduction

Heart disease is a general term for cardiovascular diseases, which itself includes related diseases with multiple sub-categories, such as cardiovascular diseases, myocardial diseases and so on. In this paper, we do not make specific distinctions and studies on each sub-category of heart disease, but generally take macroscopic heart disease as the research object. According to the research report of CDC, heart disease is the main pathological cause of death in many races, and its onset is extremely sudden and fatal. However, apart from congenital and hereditary heart disease, acquired lifestyle habits (such as smoking, drinking) are the main causes of heart disease. Based on this fact, after we have enough data and complete the relevant modeling, we can use to collect the acquired lifestyle habits of the individual and establish a personal profile, analyze, and deduce whether the individual is under the threat of heart disease. And this is the main research purpose of this paper.

This article contains multiple sections. After this Introduction is the Data section, which introduces the data itself and performs EDA analysis. After that, based on the results of the EDA analysis, this article will start modeling and model screening through Residual Analysis Plots, the model used in this article is GLM (Generalized Linear Model); the following Results section will integrate the previous research and analysis process to obtain the research results, and discuss in detail the practical significance of this article and the possible defects in the Discussion section.

Specifically, preliminary analysis of CDC-related data shows that the causes of acquired heart disease are related to a variety of influencing factors, and the most obvious ones are age, BMI, other related diseases, and so on. Interestingly, we found that heart disease rates did not differ significantly between genders, and mental health status did not significantly affect heart attack.

All the researches and analysis done in this paper are reproducible; Codes and Data can be found at Github. [1]

[1] https://github.com/Markingmark/HeartDiseases-and-Its-Cause.git

Table 1: Situation of different Age Group

Age	PhysActive	HeartDisease	Diabetes	Stroke	DiffWalk	Asthma	KidneyDisease	SkinCancer
18-24	0.8574820	0.0061717	0.0110615	0.0028959	0.0155241	0.1769370	0.0062666	0.0031808
25-29	0.8456503	0.0078443	0.0148039	0.0053082	0.0214686	0.1691536	0.0067827	0.0048953
30-34	0.8402922	0.0120514	0.0252760	0.0069855	0.0313017	0.1517091	0.0087453	0.0084786
35-39	0.8253041	0.0144039	0.0404380	0.0093917	0.0471533	0.1400973	0.0124574	0.0127981
40-44	0.8124345	0.0231362	0.0620299	0.0139484	0.0628868	0.1459107	0.0174236	0.0197087
45-49	0.7934009	0.0341425	0.0926529	0.0196411	0.0874673	0.1438667	0.0206966	0.0354734
50-54	0.7778741	0.0544874	0.1210307	0.0269876	0.1224490	0.1367899	0.0275786	0.0507840
55-59	0.7719864	0.0739994	0.1484357	0.0369997	0.1562658	0.1325066	0.0335383	0.0735289
60-64	0.7592175	0.0987651	0.1683786	0.0440242	0.1772843	0.1335273	0.0407291	0.0990916
65-69	0.7644871	0.1200843	0.1917953	0.0499253	0.1763052	0.1206407	0.0494568	0.1349302
70-74	0.7484951	0.1560277	0.2176404	0.0608724	0.1991630	0.1181716	0.0634154	0.1791405
75-79	0.7154362	0.1884834	0.2230239	0.0796015	0.2354529	0.1119076	0.0744344	0.2251653

Data and EDA

This data comes from the 2020 US CDC health questionnaire data on 400k adults(LLC 2022), The dataset contains 18 variables (9 booleans, 5 strings and 4 decimals). The data itself is intercepted from the Behavioral Risk Factor Surveillance System (BRFSS), after being sorted by CDC chosen as the research direction of this paper. The data is read and simulated by using (Wickham, Hester, and Bryan 2022) and (Wickham et al. 2019)

The core variables included in the data are: HeartDisease: Whether there is heart disease, we use 1/0 as the difference between yes/no in data processing BMI: Body Mass Index Smoking: whether to smoke AlcoholDrinking: Whether to drink alcohol Stroke: Have you ever had a stroke? PhysicalHealth: Number of days with health problems in the past 30 days MentalHealth: Number of days with mental health problems in the past thirty days DiffWalking: Is there a walking disorder Sex: Gender AgeCategory: age group

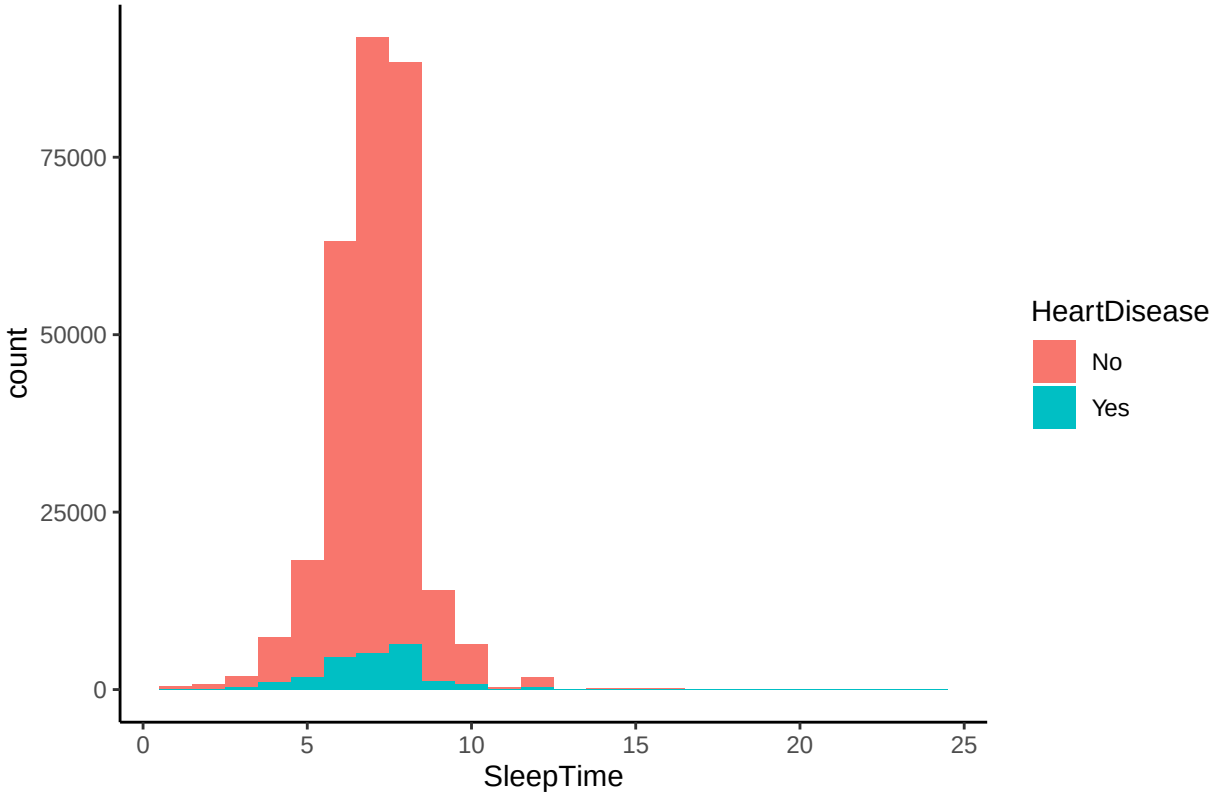
The following charts show the status of various variables in the data grouped by age. We can intuitively see that with increasing age, the amount of physical activity of individuals decreases significantly; the incidence of stroke, diabetes and other diseases increases significantly. Individuals also face higher difficulty walking as they age. At the same time, the incidence of heart disease is also positively correlated with age.

Next, we will perform a visual analysis of some data in the database to draw preliminary EDA analysis conclusions to assist the subsequent modeling process.

This graph provides a concise analysis of whether average daily sleep time is associated with heart disease by visualizing two variables, sleep time and heart disease count. As shown in the figure, the average daily sleep time of most people is mainly distributed in 7-8 hours a day, which is in line with our common sense of life. However, the changes in the graph suggest that the incidence of heart disease does not have a prominent linear relationship with sleep duration.

Although the rate of heart disease was relatively higher among those who slept less, those who slept around 5 hours a day were about twice as likely to have a heart attack as those who slept 8 hours a day, and this trend can also be seen Explain with other living habits. But we can't ignore the fact that this trend does exist, so later in the modeling process, we need to be mindful of this fact and consider adding average daily sleep time to the model.

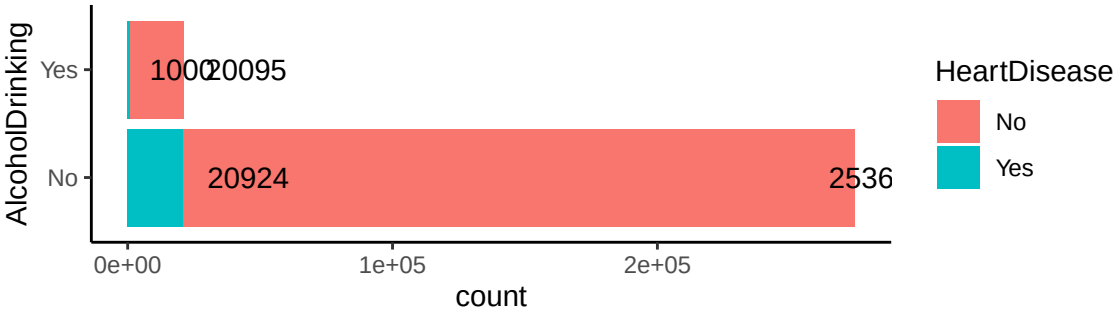
Figure 1: Histogram of Sleeptime vs Heart Disease



Indigenous succinctly visualized the link between alcohol consumption and heart disease. The vertical axis of this graph is the Categorical Variable of drinking or not, and the horizontal axis is the number of respondents. Even though the infrequent drinkers are much larger than the regular drinkers, we can still tell from the graph that the frequent drinkers have about 1.5 times the risk of heart disease than the rest of the population.

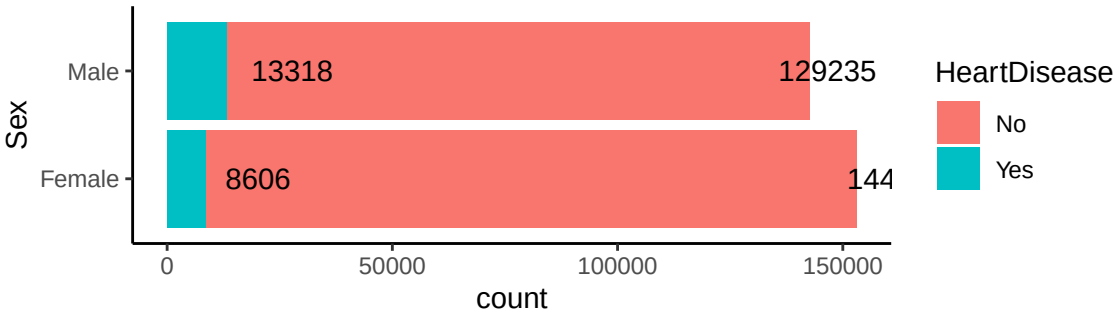
This huge difference follows the medical common sense in reality-drinking easily induces cardiovascular diseases. Based on the above facts, we will also take frequent drinking as an important influencing factor into the next model construction.

Figure 2: AlchohoDrinking and HeartDisease



The following graph (Figure 3) visualize the differences between male and female, grouping by their numbers of acquiring Heart Disease. The vertical bar shows the two different genders and if having heart disease or not is shown in the bar within the plot. As the graph shows, male has a much higher possibility of getting heart disease than females. The differences is about twice of the possibility

Figure 3: Genders and HeartDisease



Model

```
## [1] 295642
```

```
## [1] 10000
```

```
## [1] 285642
```

After data visualization and EDA, model construction in continued. The model is constructed with the help of:(**modelhelp?**) and (Simon et al. 2011) Based on the EDA and empirical analysis we first conclude the model as:

glm(Heart ~ BMI + Smoking + AlcoholDrinking + Stroke + PhysicalActivity +DiffWalking + Sex + Diabetic + AgeMid + SleepTime+ Asthma+ KidneyDisease + SkinCancer,family = binomial, data = train)

A GLM model is selected for this research as it would be easily trained and used for prediction purposes. GLM model is considerably outstanding for dealing with categorical predictors, thus it is prefect in such research purposes. The relative data about this first built model is listed below.

```
##                              BMI                      SmokingYes
##                         1.175582                         1.040335
##              AlcoholDrinkingYes                      StrokeYes
##                         1.018245                         1.043956
##             PhysicalActivityYes                    DiffWalkingYes
##                         1.158508                         1.293209
##                          SexMale                     DiabeticYes
##                         1.071588                         1.153897
## DiabeticYes (during pregnancy)                           AgeMid
##                         1.004074                         1.120341
##                        SleepTime                      AsthmaYes
##                         1.023205                         1.058357
##                KidneyDiseaseYes                   SkinCancerYes
##                         1.054936                         1.049052
```

Backward selection is then utilized by checking the AIC of model 1, listed below.

As AIC=4273.73 Comparing the AIC and deviance from predicting factors lead to the elimination of predictor:SkinCancer as it shows little correlation to the acquiring of heart disease.

```
## Start:  AIC=4273.73
## Heart ~ BMI + Smoking + AlcoholDrinking + Stroke + PhysicalActivity +
##     DiffWalking + Sex + Diabetic + AgeMid + SleepTime + Asthma +
##     KidneyDisease + SkinCancer
##
##                     Df Deviance    AIC
## - PhysicalActivity  1   4244.1 4272.1
## - AlcoholDrinking   1   4245.6 4273.6
## <none>                  4243.7 4273.7
## - SkinCancer        1   4247.4 4275.4
## - BMI               1   4248.1 4276.1
## - SleepTime         1   4255.1 4283.1
## - KidneyDisease     1   4270.7 4298.7
## - Diabetic          2   4276.9 4302.9
## - Asthma            1   4278.2 4306.2
## - Smoking           1   4280.5 4308.5
## - Stroke            1   4308.1 4336.1
## - DiffWalking       1   4310.1 4338.1
## - Sex               1   4324.5 4352.5
## - AgeMid            1   4542.1 4570.1
##
## Step:  AIC=4272.06
## Heart ~ BMI + Smoking + AlcoholDrinking + Stroke + DiffWalking +
##     Sex + Diabetic + AgeMid + SleepTime + Asthma + KidneyDisease +
##     SkinCancer
##
##                    Df Deviance    AIC
## - AlcoholDrinking  1   4245.9 4271.9
## <none>                 4244.1 4272.1
## - SkinCancer       1   4247.9 4273.9
## - BMI              1   4248.7 4274.7
## - SleepTime        1   4255.4 4281.4
## - KidneyDisease    1   4271.3 4297.3
## - Diabetic         2   4277.6 4301.6
```

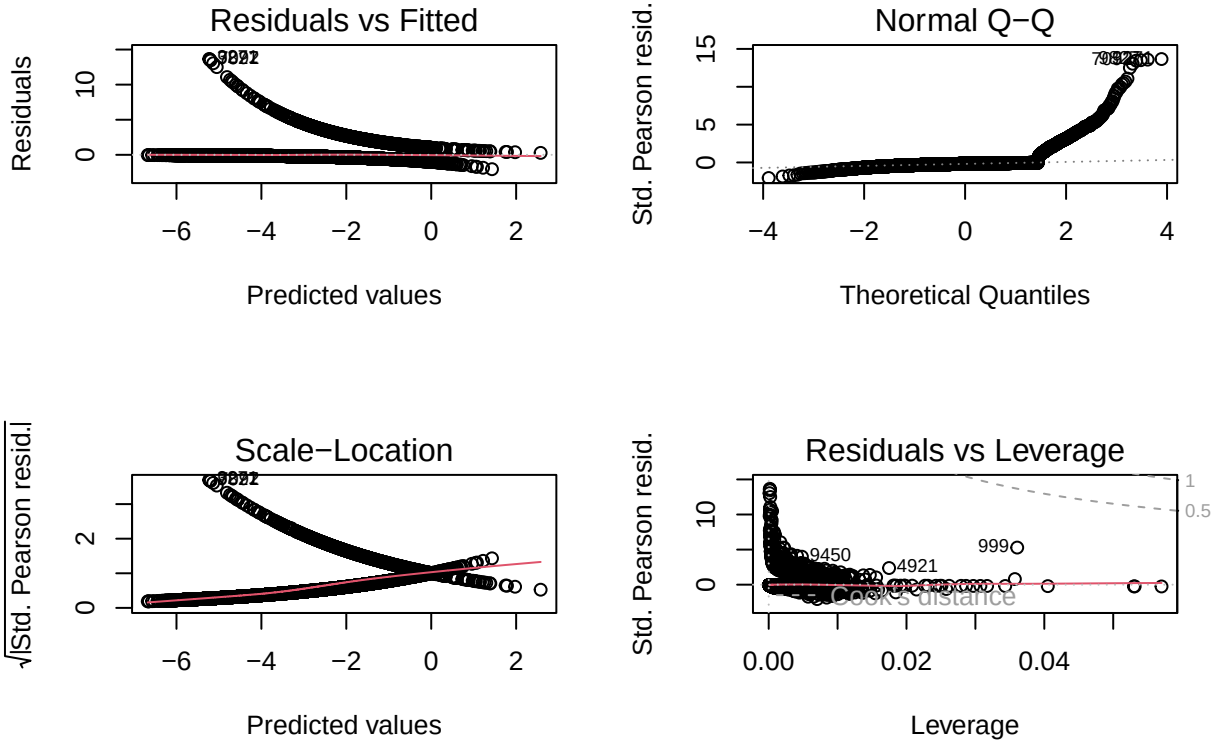
```
## - Asthma           1   4278.6 4304.6
## - Smoking          1   4281.4 4307.4
## - Stroke           1   4308.9 4334.9
## - DiffWalking      1   4317.0 4343.0
## - Sex              1   4324.6 4350.6
## - AgeMid           1   4543.1 4569.1
##
## Step:  AIC=4271.94
## Heart ~ BMI + Smoking + Stroke + DiffWalking + Sex + Diabetic +
##     AgeMid + SleepTime + Asthma + KidneyDisease + SkinCancer
##
##                  Df Deviance    AIC
## <none>                4245.9 4271.9
## - SkinCancer      1   4249.7 4273.7
## - BMI             1   4250.9 4274.9
## - SleepTime       1   4257.4 4281.4
## - KidneyDisease   1   4273.5 4297.5
## - Diabetic        2   4280.1 4302.1
## - Asthma          1   4280.5 4304.5
## - Smoking         1   4282.0 4306.0
## - Stroke          1   4311.5 4335.5
## - DiffWalking     1   4318.9 4342.9
## - Sex             1   4326.5 4350.5
## - AgeMid          1   4550.6 4574.6
```

Model 1.2 is then generated after the variable selection process.

Which is generated as: BMI + Smoking + AlcoholDrinking + Stroke + PhysicalActivity+DiffWalking + Sex + Diabetic + AgeMid + SleepTime+ Asthma+ KidneyDisease

The model's residual analytic plots are listed below. By looking at the residual analytic plots, Residual vs Fitted plots, Scale-location, and Residual vs leverage plots show no visible linear relationship between the variables, while the Normal Q-Q plot are fitted around the based line of the plot. Although the Normal Q-Q plot trends hardly prefectly, GLM model fitting is still valid considering the nature of such model.

In general, the model 1.2 fitted here is considered validate and appropriate for the research purposes. As shown below in the model summary section, most variables concluded in the model are with high signficances.



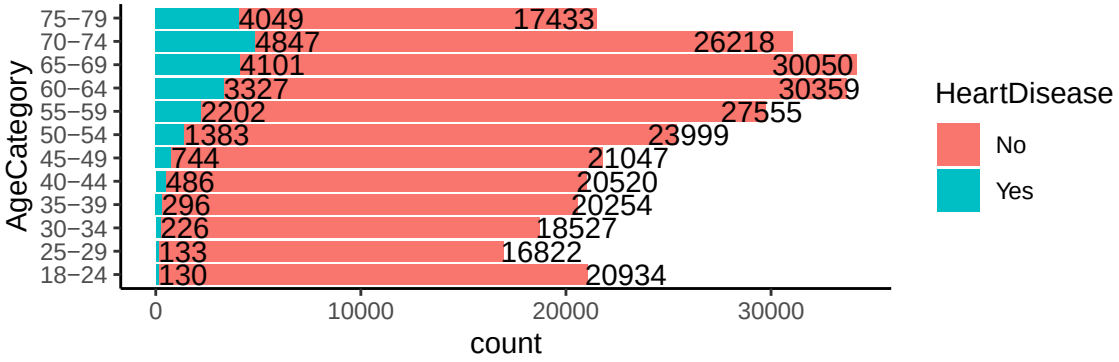
Result

The model built in this section is completed and can be used combining with the EDA sections before for conclusion purposes. Variables in this model are most considered high significance based on its summary table. Considering the residual analysis plots we investigated earlier, the model can be used to predict if an individual is facing high risk of heart disease, based on one's personal portrait from both medical and non-medical conditions.

```
##
## Call:
## glm(formula = Heart ~ BMI + Smoking + AlcoholDrinking + Stroke +
##     PhysicalActivity + DiffWalking + Sex + Diabetic + AgeMid +
##     SleepTime + Asthma + KidneyDisease, family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8142  -0.3862  -0.2465  -0.1447   3.2355
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -6.773014   0.369339 -18.338  < 2e-16 ***
## BMI                              0.014054   0.006614   2.125 0.033614 *
## SmokingYes                       0.507654   0.085149   5.962 2.49e-09 ***
## AlcoholDrinkingYes              -0.258665   0.195799  -1.321 0.186476
## StrokeYes                        1.160736   0.137932   8.415  < 2e-16 ***
## PhysicalActivityYes             -0.064726   0.096116  -0.673 0.500683
## DiffWalkingYes                   0.832986   0.100658   8.275  < 2e-16 ***
## SexMale                          0.760123   0.086339   8.804  < 2e-16 ***
## DiabeticYes                      0.561185   0.097238   5.771 7.87e-09 ***
## DiabeticYes (during pregnancy) -0.896214   1.013886  -0.884 0.376729
## AgeMid                           0.056049   0.003616  15.502  < 2e-16 ***
## SleepTime                       -0.083858   0.025175  -3.331 0.000865 ***
## AsthmaYes                        0.650323   0.107056   6.075 1.24e-09 ***
## KidneyDiseaseYes                 0.756416   0.142486   5.309 1.10e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 5322.7  on 9999  degrees of freedom
## Residual deviance: 4247.4  on 9986  degrees of freedom
## AIC: 4275.4
##
## Number of Fisher Scoring iterations: 6
```

Several conclusions can be drawn from the research. People with larger age faces higher risks of heart disease. Which is visualized and shown below in the graph. Thus, it is imperative for society to allocate more resources to pay attention to the physical health of the elderly.



On the other hand, bad lifestyles and habits leads to a significant increase of the chance of getting heart diseases. Habits include Heaving drinking, smoking, staying up late, etc. General those habits would lead to an increase of around 150%-200% of the possibilities of getting heart diseases, despising other influencing factors. Generally, it is very important to avoid bad living habits, and keep sleeping time over 7 hours is very effective at reducing the possibility of acquiring heart disease.

Discussion

Based on the personal health survey data of CDC in 2020, this paper conducts a detailed data analysis and visualization process and establishes a GLM model for data prediction and analysis. Specifically, this paper further confirms the medical factors and living habits that induce heart disease, and confirms that habits including smoking, drinking, and staying up late have an inducing effect on heart disease. The findings in this article can therefore serve as a reference for public health reports. On the other hand, this paper also confirms the vulnerability of seniors, especially senior males, of getting heart diseases. It should raise the society's attention to allocate resources to respond to this fact.

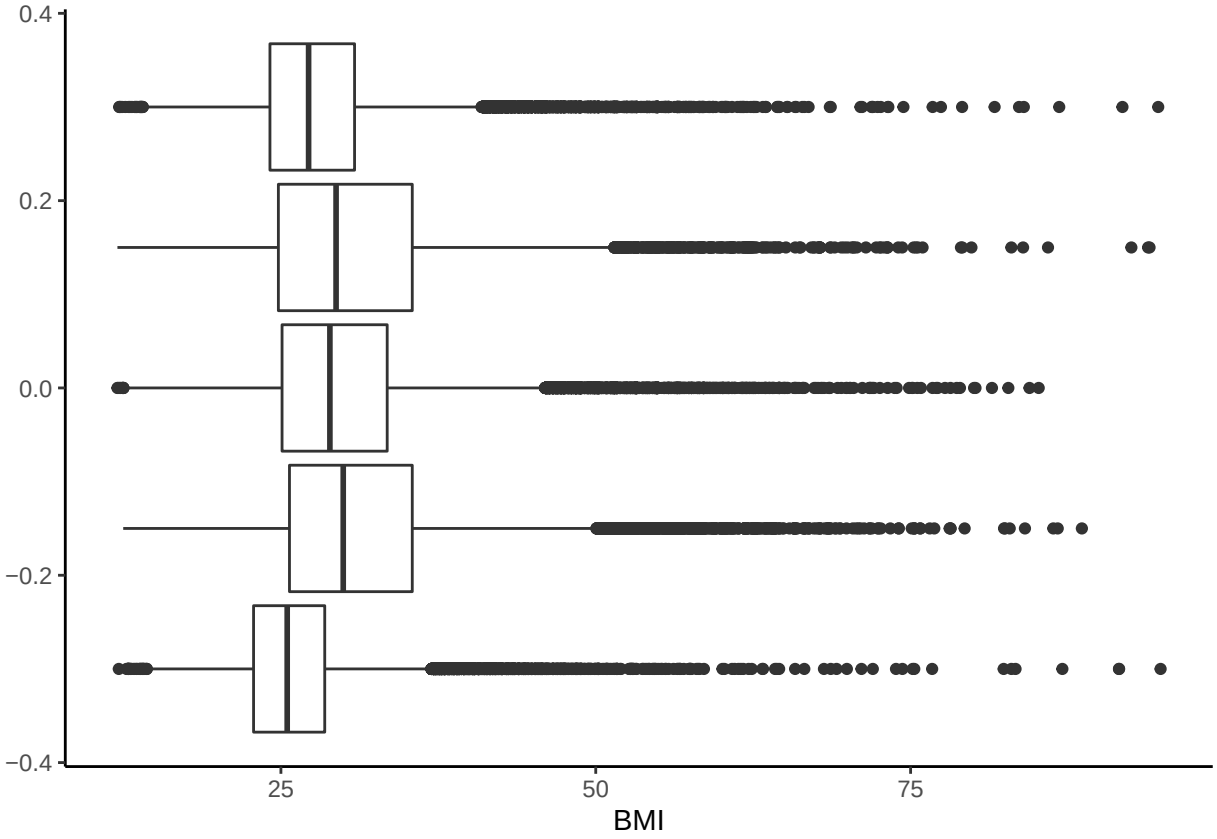
The model build within this paper can be used as a prediction tool of if one is facing high risk of heart diseases. By loading the relative and required personal information within the model, a result of 1/0, which stands for yes/no should be returned to stat if the stated risks are at large.

While the study is based on the data from CDC, the data itself should be considered at high accuracy and reliability. Thus, study based on such data should be considered reliable. While not only re-confirms the medically proven correlated factors of causing heart diseases, but this study also rules out some of the rumored factors that may cause heart diseases, such as mental status. It is shown that one's mental status has no significant impact on the chances of getting heart diseases.

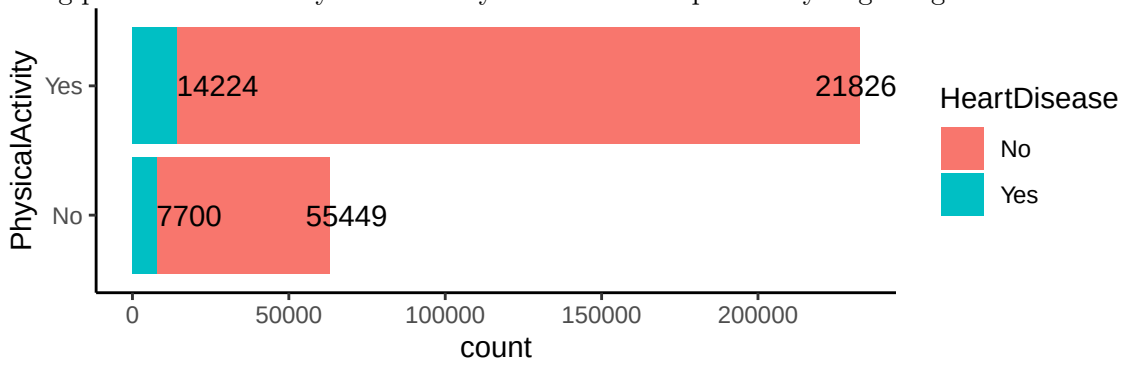
However, certain limitations do apply to this study. First, the study in done in R-Markdown with 1GB of RAM, resulting in a limited computing capability and restrained the data selection to train the model to 10000/200000+. The losing of enormous amount of data may lead to the invalidation of model, if more study is done in future. On the other hand, certain factors in the used dataset may have hidden correlation with each other, like Age and DiffWalk, BMI and Diabetes, etc. Some graphs to show the correlation effects between factors and more are listed in the Appendix section below.

Appendix

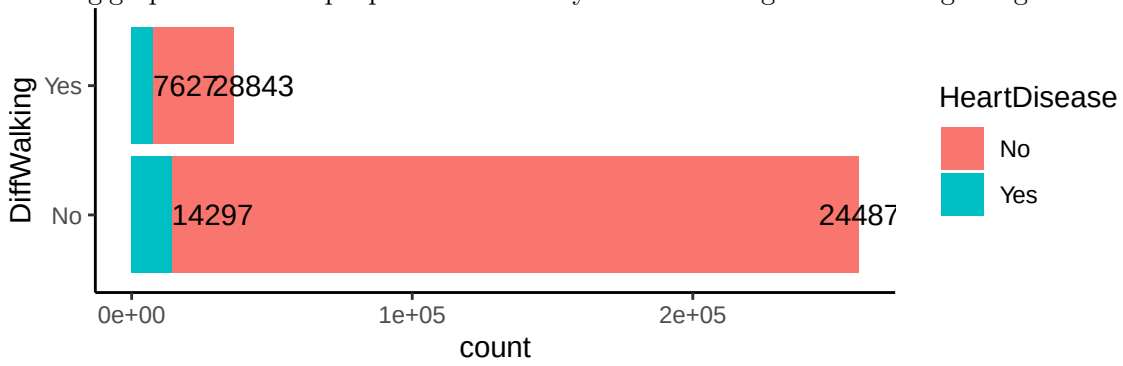
The Following boxplots show the relationship between BMI and general health. General health is reduced as BMI increases.



The following plots show that Physical Activity can reduce the possibility of getting heart disease greatly.



The following graph shows that people with disability to walk has higher chance of getting heart disease.



Reference

LLC, MultiMedia. 2022. "Personal key indicators of heart disease. Kaggle." 2022. https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Simon, Noah, Jerome Friedman, Trevor Hastie, and Rob Tibshirani. 2011. "Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent." *Journal of Statistical Software* 39 (5): 1–13. https://doi.org/10.18637/jss.v039.i05.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2022. *Readr: Read Rectangular Text Data*.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.