Effect of Frequent Vegetable Consumption on Reducing the Risk of Developing Heart Disease

Lingjun Meng - 1005712579

December 17, 2021

Abstract

Most of the population link frequent vegetable consumption to healthy eating, however it is shown in this report frequent vegetable consumption has no effect on reducing the risk of developing heart disease. Using a subset of the BRFSS 2015 dataset, with the consideration of confounding variables using Propensity Score Matching, the coefficient calculated for the treatment(eating vegetable at least one or more times per day) is not significant on a significance level of $\alpha = 0.05$. The result is generated using data collected within the U.S., results generated using different dataset collected in other countries might differ.

Keywords: Heart Disease, Eating Healthy, BRFSS 2015 dataset, Model Selection, Logistic Regression, Causal Inference, Propensity score matching.

Introduction

Background Information

As the second leading cause of death in Canada, heart disease has been a problem that affects 8.33% of Canadians aged 20 and above. [4] According to the 2012/2013 data collected by the Public Health Agency of Canada's Canadian Chronic Disease Surveillance System, there are about 12 Canadians aged 20 and above dying from heart disease every hour.[4]

Maybe the figures are just not persuasive enough, at least it is the case for me. Being a 20-years-old UofT student, I always took it for granted that diseases like heart disease are far away from me. However, that has changed since the moment I knew a friend of mine of my age was diagnosed to have a heart disease due to excessive consumption of highly processed food including burgers, sugary drinks etc. as well as being physically inactive and overweight.

Nowadays with the pandemic of the COVID-19, the life styles of every UofT students are changing. Having courses all available online made it possible for us to attend meetings in bed. That is a environment where students are able to reduce physical activities to the lowest level. [5] That is consistent with my experience, it has become my daily routine to get up before the class and then sit there for the whole day.

With the possibility of not going out for courses, there is evidence showing that a significant percentage of students are turning to junk food for their daily meal. [6] The survey collected from 100 students showed that 38% of the participants had a unhealthy change of eating habits after the transform to online courses. I believe every UofT students are familiar with the circumstance of ordering McDonald's after sitting there for a whole day's course.

Both Eating unhealthy and physical inactivity are proven to be the key risk factors of developing a heart disease.[7] The problem is that how can we reduce the risk of developing a heart disease? I have been taking a nutrition course and I learned that eating vegetables regularly can reduce the risk of developing heart disease.

However no evidence was found to support that statement. Thus, in the report, I will find out if there is a causal inference of eating vegetable everyday on the probability of developing a heart disease.

Research Question: Is there a causal inference of eating vegetables everyday on the probability of developing a heart disease?

The data used was cleaned by Alex Teboul and the original data is BRFSS2015, collected by U.S. Centers for Disease Control and Prevention in 2015. The cleaned data set contains the information including the status of the individual(whether diagnosed with heart disease) and 21 other variables that is relevant to the individual including age, sex; level of risk factors of heart diseases like smoking or not, blood pressure level and cholesterol level.

This can be associated with every UofT students and even every Canadian aged 20 and above, Now we are all living in a environment with a higher risk of developing heart disease. Not to mention other causes of heart disease like smoking. After reading this report, hopefully students will pay more attention to heart disease and keep a healthy lifestyle in this difficult time.

Terminologies Used

Heart Disease

This is a term that refers to a series of different type of heart situations, the most common heart disease in United States is coronary artery disease. [8] Coronary artery disease will usually cause the blood low to be slower, which can be the reason of a heart attack. The binary target variable is defined as whether diagnosed with heart disease or ever had a heart attack.

Cholesterol Level

Total Cholesterol: The total amount of blood cholesterol in your body, consisting of HDL cholesterol, LDL cholesterol and other lipids. Lower values of Total Cholesterol are considered to be healthier. A number over 240 mg/dL[9] is considered to be high for adults. It is usually used as a health indicator for heart disease and this is the factor used by the dataset I will be using.

HDL Cholesterol: The abbreviation for high-density lipoprotein cholesterol, also known as "good cholesterol". Meaning that a higher number of HDL Cholesterol can lower the risk of developing a heart disease.[10]

LDL Cholesterol: The abbreviation for low-density lipoprotein cholesterol, also known as "bad cholesterol". Meaning that a higher number of LDL Cholesterol can increase the risk of developing a heart disease by building up walls in your arteries.[10]

Blood Pressure

Similar as Cholesterol Level, Blood Pressure is also a risk indicator for heart disease. [8] The higher the blood pressure, the higher risk of developing a heart disease. This is also included in the dataset. Individuals will be diagnosed to have high blood pressure if systolic blood pressure is 130 mm Hg or higher and/or diastolic blood pressure is 80 mm Hg or higher.[11]

Hypothesis

Null hypothesis: Vegetable consumption does not have a causal inference on the probability of developing heart disease.

Alternative hypothesis: Vegetable consumption has a causal inference on the probability of developing heart disease.

Data

Data Collection Process

The data [12] I will be using is a cleaned dataset published on Kaggle, the data set was generated by selecting specific features from the BRFSS2015 dataset [13], which is also available on Kaggle.

The original dataset (BRFSS2015) was collected by Centers for Disease Control and Prevention of U.S. through a telephone survey on U.S. residents in 2015. During the survey, residents' information about health-related risk behaviors as well as chronic health conditions were collected. [14]

The data I will be using is a subset of the original BRFSS2015 data, Alex Teboul selected 22 variables from the original 330 variables. The selected variables include the information like having heart disease or not, situation of drinking, smoking, eating vegetables as well as some general information including age, sex, etc. Since the dataset is already cleaned, there are no missing values in each of the observations.

Data Summary

The dataset have 253680 observations with 22 variables, all the variables were transformed into numerical variables. The categories of variables that are originally categorical are replace with integer 0,1,2, etc., representing different classes.

No further modification was made to the dataset cleaned by Alex Teboul, he also has a page introducing the specific and detailed cleaning process of obtaining the dataset from the original BRFSS 2015 data. Please refer to reference #12 for more details.

Below is a table summarizing all the variables in the raw dataset with descriptions. The true numerical variables(not transformed from categorical variables)'s mean and standard deviation are represented in the format of *Mean(SD)*. While the definition of the classes of the variables that is originally categorical is explained. For those who are interested in the raw data, please refer to Appendix for a glimpse of the raw dataset.

Table 1: Summaries of all the variables in the raw dataset

Variables	Descriptions	Summaries
MentHlth	Number of days respondent had problems related to mental health in the past month	3.1847722(7.4128467)
PhysHlth	Number of days respondent had problems related to physical health in the past month	4.2420806(8.7179513)
BMI	Body Mass Index of respondent	28.3823636(6.6086942)
HighBP	Whether respondent have High Blood Pressure	0=No, 1=Yes
HighChol	Whether respondent have High Cholesterol Level	0=No, 1=Yes
CholCheck	Whether respondent have Cholesterol check within last 5 years	0=No, 1=Yes
HeartDiseaseorAttack	Whether respondent have Heart Disease	0=No, 1=Yes
Smoker	Whether respondent smoked at least 100 cigarettes in the entire life	0=No, 1=Yes
Stroke	Whether respondent ever told to have a stroke	0=No, 1=Yes
Diabetes	Whether respondent have Diabetes	0=No, 1=Pre-Diabetes, 2=Diabetes
PhysActivity	Whether respondent exercised in the past 30 days	0=No, 1=Yes

Variables	Descriptions	Summaries
Fruits	Whether respondent consumes fruit 1 or more times per day	0=No, 1=Yes
Veggies	Whether respondent consumes vegetables 1 or more times per day	0=No, 1=Yes
HvyAlcoholConsump	Whether respondent have more than 14 drinks a week for men and 7 for women	0=No, 1=Yes
AnyHealthcare	Whether respondent have health coverage	0=No, 1=Yes
NoDocbcCost	Whether respondent could not see a doctor due to cost in the past year	0=No, 1=Yes
GenHlth	How would respondent rate their general health	1=most unhealthy and 5=very healthy
DiffWalk	Whether respondent have difficulties walking or climbing stairs	0=No, 1=Yes
Sex	Indicating respondent's Sex	0=Female, 1=Male
Age	Indicating respondent's Age group	1=Youngest, 13=Oldest
Education	Indicating respondent's Educational Level	1=Lowest, 6=Highest
Income	Indicating respondent's Income Level	1=Lowest, 8=Highest

Since all the missing values have already been removed from the dataset, the dataset is already cleaned. Here the target variables is "Heart Disease or Attack" variable, which is a binary variable stating whether this individual has been diagnosed with Heart Disease or ever had Heart Attack. 0 represents no while 1 represents yes. Among the 253680 observations, 23893 of them are diagnosed with Heart Disease or ever had Heart Attack, below is a graph visualizing the distribution.

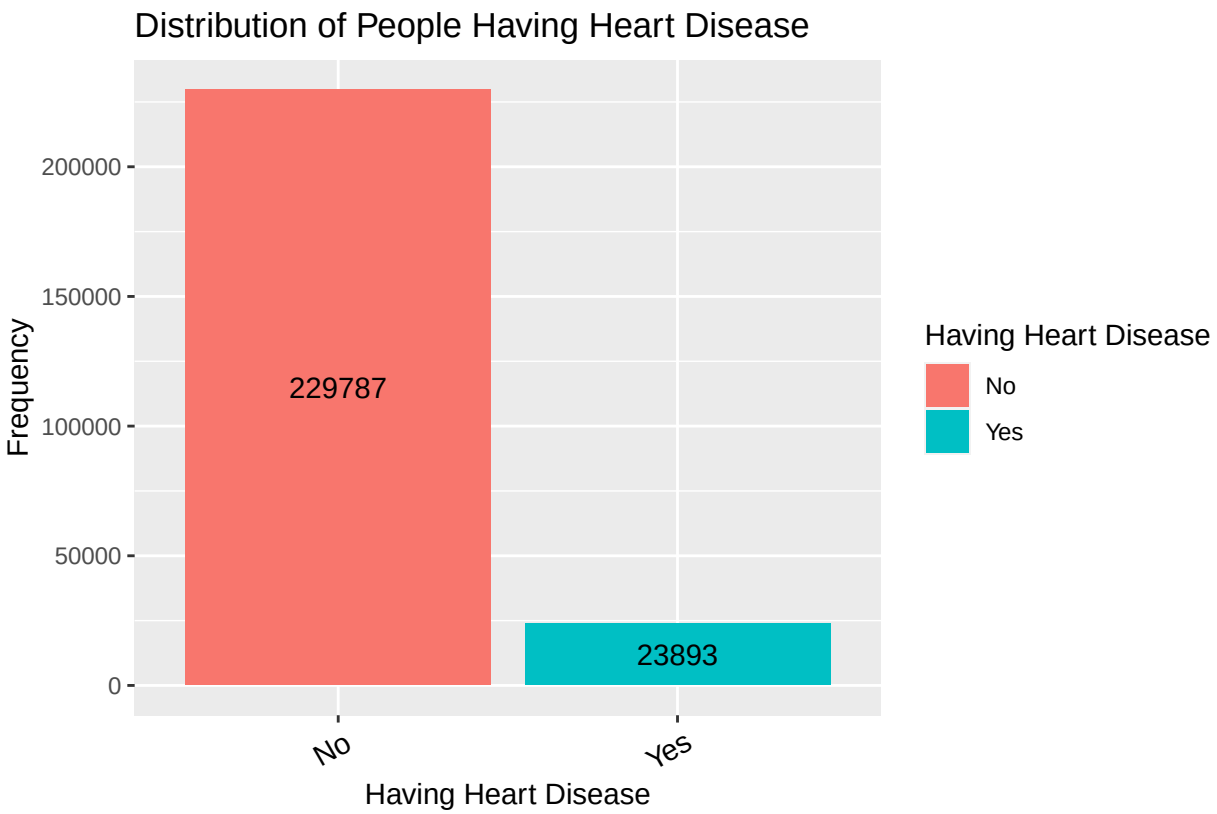

Figure 1. Distribution of Individuals Diagnosed with Heart Disease

From Figure 1, the proportion of the respondents that were diagnosed with Heart Disease is at approximately 0.094. That means on average there is one individual diagnosed with Heart Disease in every 11 individuals.

The variable that also contributes to the goal is the "Veggies" variable. It is defined as whether an individual consumes vegetables 1 or more times per day. Similar to the binary target variable, 0 stands for no while 1 stands for yes. This is also the treatment variable to carry out a Propensity Score Matching. Among all the observations in the data set, approximately 0.812(81.2%) of the 253680 individuals consume vegetables 1 or more times every day.

Below is a graph visualizing the relationship between these two binary variables.

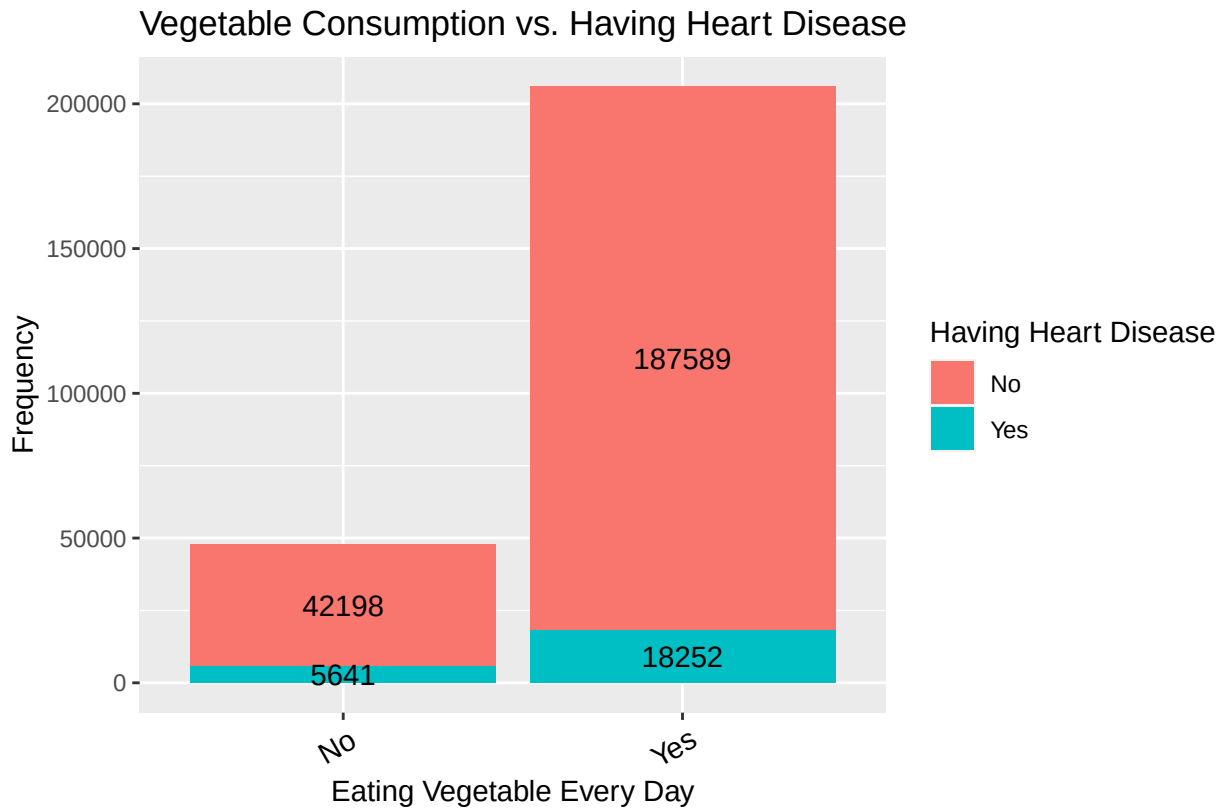

Figure 2. Relationship Between Vegetable Consumption and Having Heart Disease

From Figure 2, it is intuitive that among all the respondents, individuals who consume vegetables every day have a slightly lower proportion of people who have Heart Disease. The proportion for the Eating-Vegetable-Everyday group is 0.089, while this figure for the other group(Not Eating Vegetable Everyday) is 0.118.

All analysis for this report was programmed using `R version 4.0.5`.

Methods

Model Selection

This is a general statistical method used to choose the best model from all the available ones. In this report, Model selection method is used to select the best model that will be used for Propensity Score Matching.

Since there are 20 variables left except the response variable " Heart Disease or Attack" and the treatment variable "Veggies", it will be extremely time-consuming to generate the model manually. A practical way of solving this problem is usually using Automated Selection method.

Automated Selection Method

There are many Automated Selection Method that is available for generating the Propensity Score Matching Model and different Method have different advantages and disadvantages. Here the model we want is used to carry out PSM, the only requirement of the model is to be accurate at calculating Propensity Scores and there is no requirement on the model complexity.

Stepwise Selection Using AIC Penalty :

AIC stands for Akaike Information Criterion, which is a measure that tells us how well a model performs. This measure is chosen as it punishes less on the complexity of the model and that fits our goal to make the model accurate. Generally the smaller the AIC the better.[15]

Depending whether we will start with a full model(including all predictors) or a null model(with 0 predictors in the model), the Stepwise Selection method will work differently. If we choose to start with a full model with P predictors, then the first step will be a elimination step where the program will calculate the AIC values for all the possible models with *P-1* predictors. If the smallest AIC for the smaller model is smaller that the previous model, then the smaller model with smallest AIC value will be chosen to go to the next addition step. Adding a predictor is similar with elimination, AIC values for all the possible models with one more predictor will be calculate and the smallest one will be selected. And then the program will iterate between addition and elimination until we can not add or remove anymore.

The reason why Stepwise Selection method is used instead of Forward or Backward is obvious, as Stepwise method takes the conditional nature of regression in to consideration. Thus, the algorithm will check and make sure to include the predictors that explains a significant portion of the variation regardless the changing conditional relationship between the treatment variable and the p predictors.[16]

There are assumptions of using Automated Selection methods, we have to assume the model does not have multicollinearity problems. This is an important assumption to check as the algorithm will still work even the model have multicollinearity problems. One possible way of checking multicollinearity is using the VIF factor. VIF stands for Variance inflation factor, a cutoff of 5 is usually adopted.[17] The most commonly acceptable way of dealing multicollinearity is to remove at least one variable in the model that has a VIF bigger than 5, until all the variables in the model have a VIF smaller than 5.

Still, there are drawbacks of using automated selection method, please refer to limitations section for more information.

Propensity Score Matching

PSM is a statistical method used to measure the effect of treatment on the outcome with the consideration of all the confounding variables. Here the treatment refers to frequent vegetable consumption and the outcome is the status of heart disease. Confounding variables here refers to all other variables that may also affect the outcome, for example the blood pressure and the cholesterol level.

To be more specific, we can not simply conclude that eating vegetable will lower the risk of developing heart disease without checking whether there are other variables that may potentially affect the risk of developing heart disease. For example, There could be a possibility that people consume vegetable on a daily basis also have a relatively lower cholesterol level or blood pressure.

After a Propensity Score model is generated, for every observation, both in the treatment group and control group, there will be a Propensity Score calculated. Then for each observation in the treatment group, there will be one observation in the control group having the closest Propensity Score matched to it. This is also known as nearest neighbor matching. There are other matching method available but here we will be using this method as it is simple to carry out.

Here in our situation, the propensity scores will be matched for each individual who does not consume vegetable on a daily basis with one individual that consumes vegetable on a daily basis.

Then using the paired observation to find if causal inference exists will minimize the effect that all other

confounding variables brought to the response variable. However, PSM is not perfect, please see limitations section for the disadvantages of applying PSM.

One important assumption of PSM is the randomization of the treatment. That requires the dataset where the treatment variable comes from to be completely observational data instead of experimental data. Here in our case, this assumption is satisfied as the data is collected through phone survey. And the treatment variable here is consumption of vegetable, which will be randomly assigned using the propensity scores calculated.

Logistic Regression

This is a statistical method used to model a binary outcome using one or more predictors. Here after creating paired observations using propensity score matching for the treatment variable, I will be using a Logistic Regression model to model the status of the individual using the treatment predictor where the effect all confounding variables are minimized. That is, examine the effect of frequent vegetable consumption on the risk of developing a heart disease.

Logistic Model

$$\log(P/(1-P)) = \beta_0 + \beta_1 X_{Treatment} + \epsilon$$

Here P refers to the probability of having a heart disease;

β_0 refers to the intercept;

β_1 refers to the coefficients for being in the Treatment group;

$X_{Treatment}$ refers to the status in the treatment group, if the observation is in the treatment group, $X_{Treatment} = 1$, otherwise $X_{Treatment} = 0$;

ϵ refers to the error factor.

Note that the $X_{Treatment}$ variable here is different from the "Veggies" variable in the raw dataset, instead, it is matched(paired) observations by their propensity scores predicted using the Propensity Score Model.

In this particular logistic regression model, only one binary predictor is included. Thus the coefficients β_0 and β_1 have special meanings here. β_0 here also refers to the expected log probability($\log(P/(1-P))$) of the control group while $\beta_0 + \beta_1$ refers to the expected log probability of the treatment group.

The output of the regression model is in the form of log probability, but what we want is actually the exact probabilities. One way that is usually considered to convert the log probabilities to common probabilities is the Sigmoid function.

Sigmoid Function

$$\sigma(x) = 1/(1+e^{-x})$$

Here x refers to the log probabilities;

e refers to the base of the natural logarithm.

Then $\sigma(x)$ will be the probability we want.

Results

Propensity Model

Through the Automated Selection Method: Stepwise Selection using AIC penalty, only variables "Cholesterol check", "Diabetes","AnyHealthcare" and "DiffWalk" were removed and the rest are all included in the Propensity Model(except the target variable of having heart disease).

On a significance level of $\alpha = 0.05$, all the predictors included in the propensity model are significant. Then the model is used to predict the propensity score for each of the observation in order to create matches. Please see Appendix for a summary of the Propensity Model, including estimated coefficients for the predictors as well as their significance level.

Propensity Score Matching

After the propensity scores were predicted for each observation, each individual in the control group(not eating vegetables everyday) will be paired with an individual in the treatment group(eating vegetables everyday). It is not the other way around because the observations in the control group are much fewer than that in the treatment group. Note that in Data section, about 81.2% percent of the 253680 observations are in the treatment group and rest in the control group.

The observations in the treatment and control group with the closest predicted propensity scores are paired, this makes the treatment and control as similar as possible(of course, except the treatment variable). All the observations in the treatment group that have not been matched to a observation in the control group were removed. There are 95678 observations left in the matched dataset, 47839(half) of them in the treatment group and rest in the control group.

Logistic Model

The logistic model of having heart disease is generated with only the treatment predictor: consumption of vegetables using the matched dataset. No other predictors are included since we have already accounted for the confounding variables when generating the propensity model. Note that almost all the variables were included in the propensity model, so only the treatment variable is included in the final logistic model for simplicity. Please see the following table for the transformed summary of the logistic model.

Table 2: Summary of the Final Logistic Model

Coefficient	estimates	P-value	Significance
β_0	-2.01231	<2e-16	Extremely Significant
β_1	0.01955	0.328	Not Significant

Since the coefficients of logistic model are in the form of log probabilities, Sigmoid function was used to transform these log probabilities into common probabilities. Connecting Table 2 with what has been discussed above, Here $\sigma(\beta_0)$ refers to the probability of developing heart disease in the control group while $\sigma(\beta_0 + \beta_1)$ refers to the probability of developing heart disease in the treatment group.

Among the matched pairs, the probability of individuals in the treatment to develop a heart disease is 0.12, while this figure for the control group is 0.118. The statistics and the P-values calculated showed that with the considerations of confounding variable, consuming vegetables everyday does not have an impact on the risk of developing a heart disease.

Conclusions

The goal of the report is to find if there is a causal inference of eating vegetables everyday on the risk of developing a heart disease. The null hypothesis sates that frequent vegetable consumption has no impact on the risk of developing a heart disease while alternative hypothesis says there is a causal inference between frequent vegetable consumption and the probability of developing a heart disease.

Automated Model Selection method was adopted to generate the logistic model used for predicting the propensity scores of being in treatment group. Then treatment and control observations were matched using Propensity Score Matching method and observations that have not been paired were removed.

A logistic regression model was then generated using the dataset that only has paired observations to find out the effect of the treatment on the response variable. The response binary variable is whether the individual is diagnosed with heart disease or ever had heart attack, and the treatment variable is whether the individual consumes vegetables everyday.

From the final logistic model generated, the P-value of coefficient calculated for the treatment variable is 0.328. On a significance level of $\alpha = 0.05$, the impact of the treatment on the response variable in not significant.

Key Findings

Even the probability of developing a heart disease is lower for people that consume vegetables everyday from the raw dataset, frequent vegetable consumption has no impact on the risk of developing heart disease with the consideration of confounding variables. That is *Frequent Vegetable Consumption has no Causal Inference on the Risk of Developing Heart Disease.*

This is not a surprising finding, note that here Frequent vegetable consumption is defined as eating vegetable one or more times per day. The result does not suggest that we do not have to eat healthy. While eating healthy has been proven to reduce the risk of developing heart disease,[18] the treatment variable is slightly different with eating healthy. There could be many possible reasons why there is no causal inference between the treatment and the response, but the results did suggest that just eating vegetables everyday is not sufficient to reduce the risk of developing a heart disease.

Limitations

1. Although the matched dataset have 95678 observations after PSM, the cleaned dataset is only a subset of the original BRFSS 2015 dataset. There were originally 441,456 observations , and there were only 253,680 observations left after the data is cleaned by Alex Teboul.[12] Thus, data used may be less representative than the original dataset.
2. The dataset comes from the BRFSS 2015 dataset, which is about 6 years ago. The newest BRFSS data is released on 2020. The newest dataset may be more representative for the population than the dataset used.
3. The propensity model generated using Automated Selection Method might not be the best model available as the function did not went through all the possible models that is available.
4. Even Propensity Score Matching is easy to perform and wide used, it is shown that PSM would usually do the opposite of its intended goal.[19] It is proved that rather than minimizing the imbalance between treatment and control group, PSM usually increase the imbalance and thus increase bias.[19]
5. The model and the analysis was made based on only the observations in the United States. Thus, there could be a different result if the data comes from other countries in the world.
6. Even the assumptions was validated for the propensity model, model diagnostics of influential points was omitted. Although it is not necessary, but doing the model diagnostics can indeed make the propensity model more accurate.

Next Steps

As discussed in the Limitations, researchers interested in this topic can use the BRFSS 2020 data instead to carry out the study. Can manually compare all the models possible if time permits, this guarantees the propensity model to be the best one available.

Also include a part of model diagnostics of influential observations of the propensity model to make the propensity scores predicted more accurate. In addition, other Matching Methods can be considered when trying to match each treatment observation with control ones.

If the result still shows there is no causal inference of frequent vegetable consumption on the risk of developing heart disease. there are still other topics available. For instance effects of frequent fruit consumption, nuts consumption, etc. on reducing the risk of developing heart disease are also interesting.

Discussion

This report showed that merely frequent vegetable consumption is not sufficient to reduce the risk of developing heart disease. I believe most of the population have linked frequent vegetable consumption to healthy eating, however, it turns out to be false. This again reminds us the importance of eating healthy, not only should people have regular vegetable consumption, but also pay attention to unhealthy foods.

Under the pandemic that COVID-19 brought to us, every individual should pay extra attention to their health status. A healthy body together with a healthy mind will be substantial to any academic and/or career success.

Bibliography

1. Grolemund, G. (2014, July 16) *Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/articles_intro.html. (Last Accessed: January 15, 2021)
2. Dekking, F. M., et al. (2005) *A Modern Introduction to Probability and Statistics: Understanding why and how.* Springer Science & Business Media.
3. Allaire, J.J., et. el. *References: Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/docs/. (Last Accessed: January 15, 2021)
4. Canada.ca. 2017. Heart Disease in Canada. [online] Available at: https://www.canada.ca/en/public-health/services/publications/diseases-conditions/heart-disease-canada.html [Accessed 7 December 2021].
5. BALRAM, A., 2020. How online learning can affect student health. [online] The Johns Hopkins News-Letter. Available at: https://www.jhunewsletter.com/article/2020/04/how-online-learning-can-affect-student-health [Accessed 7 December 2021].
6. NACHBAR, Q., 2020. Online school impacts eating habits of CHS students. [online] THE SANDPIPER. Available at: https://thesandpiper.org/online-school-impacts-eating-habits-of-chs-students/ [Accessed 7 December 2021].
7. Centers for Disease Control and Prevention. 2021. Know Your Risk for Heart Disease [online] Available at: https://www.cdc.gov/heartdisease/risk_factors.htm [Accessed 16 December 2021].
8. Centers for Disease Control and Prevention. 2021. Heart Disease Resources | cdc.gov. [online] Available at: https://www.cdc.gov/heartdisease/about.htm [Accessed 16 December 2021].
9. Medicalnewstoday.com. 2021. Cholesterol levels by age: Differences and recommendations. [online] Available at: https://www.medicalnewstoday.com/articles/315900 [Accessed 16 December 2021].
10. WebMD. 2020. Understanding Cholesterol Numbers. [online] Available at: https://www.webmd.com/cholesterol-management/guide/understanding-numbers#091e9c5e800081fd-2-6 [Accessed 16 December 2021].
11. Centers for Disease Control and Prevention. 2021. High Blood Pressure Symptoms, Causes, and Problems | cdc.gov. [online] Available at: https://www.cdc.gov/bloodpressure/about.htm [Accessed 16 December 2021].
12. Teboul, A., 2021. Heart Disease Health Indicators Dataset Notebook. [dataset] Kaggle.com. Available at: https://www.kaggle.com/alexteboul/heart-disease-health-indicators-dataset-notebook [Accessed 3 December 2021].
13. Centers for Disease Control and Prevention. 2017. Behavioral Risk Factor Surveillance System. [dataset] Available at: https://www.kaggle.com/cdc/behavioral-risk-factor-surveillance-system [Accessed 3 December 2021].
14. cdc.gov. 2021. CDC - BRFSS - BRFSS Frequently Asked Questions (FAQs). [online] Available at: https://www.cdc.gov/brfss/about/brfss_faq.htm [Accessed 3 December 2021].
15. En.wikipedia.org. 2021. Akaike information criterion - Wikipedia. [online] Available at: https://en.wikipedia.org/wiki/Akaike_information_criterion [Accessed 16 December 2021].
16. En.wikipedia.org. 2021. Stepwise regression - Wikipedia. [online] Available at: https://en.wikipedia.org/wiki/Stepwise_regression [Accessed 16 December 2021].
17. En.wikipedia.org. 2021. Variance inflation factor - Wikipedia. [online] Available at: https://en.wikipedia.org/wiki/Variance_inflation_factor [Accessed 16 December 2021].
18. Medicalnewstoday.com. 2021. The top 10 benefits of eating healthy. [online] Available at: https://www.medicalnewstoday.com/articles/322268 [Accessed 17 December 2021].

19. King, G. and Nielsen, R., 2019. Why Propensity Scores Should Not Be Used for Matching. Political Analysis, 27(4), pp.435-454.

20. Mine Çetinkaya-Rundel et al, 2021, openintro: Data Sets and Supplemental Functions from 'OpenIntro' Textbooks and Labs, R package version 2.2.0.

21. Hadley Wickham, 2021, tidyverse: Easily Install and Load the 'Tidyverse', R package version 1.3.1.

22. Yihui Xie, 2021, knitr: A General-Purpose Package for Dynamic Report Generation in R, R package version 1.34.

23. Brian Ripley et al, 2021, MASS: Support Functions and Datasets for Venables and Ripley's MASS, R package version 7.3-5.

24. John Fox et al, 2021, car: Companion to Applied Regression, R package version 3.0-12.

25. Andrew Gelman et al, 2021, arm: Data Analysis Using Regression and Multilevel/Hierarchical Models, R package version 1.12-2.

Appendix

A1: Ethics Statement

Reproducibility Considerations

In the model selection part, the Automated Selection method of Stepwise Selection using AIC penalty can easily get crashed when the model is logistic regression. Or more specifically, the *stepAIC* function in R will often crash when using Rstudio with Jupyter hub. This is because there are too many observations(253680) and the Jupyter hub server can easily be dead. This can be a reproducibilty problem to those who are using Rstudio with Jupyter hub.

In addition, the *matching* function in arm package can take a very long time to run as there were 4.7839×10^4 pairs to match. No movement can be done when the function is running otherwise the program will crash. This is inevitable but yet can still cause reproducibility problems.

Avoid P-hacking

Note that it is the most important thing for people in the statistics field to be honest with the research we are doing. Here in my example, even with the intuition that vegetable consumption would reduced the risk of developing heart disease, I was still honest with the result generated. The probability of developing heart disease for people frequently consume vegetables is even slightly higher than that of people do not frequently consume vegetables. This is astonishing but understandable after a thorough consideration.

Of course we have to make sure there is nothing wrong with the method, however, if the method itself is correct, we should not manipulate the data to get what we want. We have to respect the truth and that is usually how scientific breakthrough occurs.

A2: Materials

Glipse of the Raw Dataset

```
## Rows: 253,680
## Columns: 22
## $ HeartDiseaseorAttack <int> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0~
## $ HighBP               <int> 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 0, 1, 0, 1, 1~
## $ HighChol             <int> 1, 0, 1, 0, 1, 1, 0, 1, 1, 0, 0, 1, 0, 1, 1, 0, 1~
## $ CholCheck            <int> 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ BMI                  <int> 40, 25, 28, 27, 24, 25, 30, 25, 30, 24, 25, 34, 2~
## $ Smoker               <int> 1, 1, 0, 0, 0, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 0, 0~
## $ Stroke               <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0~
## $ Diabetes             <int> 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 2, 0, 0, 2, 0, 0, 0~
## $ PhysActivity         <int> 0, 1, 0, 1, 1, 1, 0, 1, 0, 0, 1, 0, 0, 0, 1, 1, 1~
## $ Fruits               <int> 0, 0, 1, 1, 1, 1, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0, 1~
## $ Veggies              <int> 1, 0, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1~
## $ HvyAlcoholConsump    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ AnyHealthcare        <int> 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ NoDocbcCost          <int> 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0~
## $ GenHlth              <int> 5, 3, 5, 2, 2, 2, 3, 3, 5, 2, 3, 3, 3, 4, 4, 2, 3~
## $ MentHlth             <int> 18, 0, 30, 0, 3, 0, 0, 0, 30, 0, 0, 0, 0, 0, 30, ~
## $ PhysHlth             <int> 15, 0, 30, 0, 0, 2, 14, 0, 30, 0, 0, 30, 15, 0, 2~
## $ DiffWalk             <int> 1, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 1, 0, 1, 0, 0, 0~
## $ Sex                  <int> 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0~
## $ Age                  <int> 9, 7, 9, 11, 11, 10, 9, 11, 9, 8, 13, 10, 7, 11, ~
## $ Education            <int> 4, 6, 4, 3, 5, 6, 6, 4, 5, 4, 6, 5, 5, 4, 6, 6, 4~
## $ Income               <int> 3, 1, 8, 6, 4, 8, 7, 4, 1, 3, 8, 1, 7, 6, 2, 8, 3~
```

A3: Supplementary Plots

Summary of the Propensity Model

Table 3: Summary of the Full Propensity Model

Predictors	Estimate	Std. Error	P-value	Significance
(Intercept)	-0.7089230	0.0464993	< 2e-16	***
HighBP	-0.0479829	0.0123264	9.91e-05	***
HighChol	-0.0300317	0.0116301	0.00982	**
BMI	-0.0022383	0.0008209	0.00640	**
Smoker	0.1377811	0.0112409	< 2e-16	***
Stroke	-0.1416940	0.0251387	1.74e-08	***
PhysActivity	0.4911854	0.0120347	< 2e-16	***
Fruits	1.1692827	0.0109632	< 2e-16	***
HvyAlcoholConsump	0.2526824	0.0251590	< 2e-16	***
NoDocbcCost	0.0879502	0.0190636	3.96e-06	***
GenHlth	-0.0843585	0.0065054	< 2e-16	***
MentHlth	-0.0033861	0.0007424	5.10e-06	***
PhysHlth	0.0052497	0.0007093	1.35e-13	***
Sex	-0.3552623	0.0110843	< 2e-16	***
Age	0.0049252	0.0019988	0.01373	*
Education	0.1902531	0.0058934	< 2e-16	***
Income	0.1040960	0.0029326	< 2e-16	***