DATA 102 Final Project

Amber Shao, Iris Wu,
Sravan Mokkala, William Vereyken

University of California, Berkeley
12/12/2022

Contents

1 Data Overview

The data is a 2020 annual CDC (Center for Disease Control and Prevention) survey data of 400k adults related to their health status. It is a major part of the Behavioral Risk Factor Surveillance System (BRFSS), which "conducts annual telephone surveys to gather data on the health status of U.S. residents". The original dataset had 401,958 rows and 279 columns, with each row representing an individual, and each column representing the answers to the survey questions about one's health status, such as "Do you have serious difficulty walking or climbing stairs?". The data has subsequently been cleaned such that there were only 18 most relevant and prominent features that are diseases and conditions that directly impact one's health status. The data that we chose comes from a public dataset published on Kaggle.

We did not choose additional data sources. The dataset we obtained contained an adequate and appropriate amount of information for us to conduct our research. Our data represents a sample of 400,000 adults taken from the overall adult population of the United States. The sample aims to pick from all states and regions to mimic the overall population distribution. The key demographic variables being tracked in the sample are sex, age, and race. The sample is 48% male and 52% female which is very close to the near 50-50 split in the entire country. The age and race distributions of the sample are also very similar to the overall population. Both have 16% of adults between the ages of 65 and 75 (the largest age group) and around 10% of adults in both the 18 to 24 and 24 to 34 age ranges. 75% of both the sample and the population are white, with minorities following a similar pattern. There are no noticeable differences between the key metrics of the dataset and the overall population and this is most likely because the sample was conducted by the Center for Disease Control and Prevention (CDC), which has access to census data and aims to accurately represent the overall population to ensure that their research is translatable. This points to a strong degree of generalizability of our results.

According to the BRFSS website, having all information publicly and freely available, it was indicated that the data was used for helping "establish and track state and local health objectives, plan health programs, implement disease prevention and health promotion activities, and monitor trends. Nearly two-thirds of states use BRFSS data to support health-related legislative efforts." The data was conducted via telephone interviews, and on the data collection related pages, it was indicated that "Reports from these surveys never contain any personal information and never identify who participated in the survey.", Privacy is required by law. Names and all other personal identifiers will not be released. The facts collected in the surveys will only be provided in summary reports." Moreover, interviewers may choose to opt-out of the interview process if they wish to do so.

The dataset is quite granular as each row represents an individual's response (thus the granularity is per person). This means that our interpretation of our findings for the causal effect questions is at the individual level - e.g. a particular character has a causal effect on individuals. For the hypothesis testing, we are looking for differences in population means so the granularity is less important for the interpretation; however, the higher sample size makes the underlying statistics more certain and allows us to reject the null at higher confidence levels (as compared to a sample size that is far too small creating an underpowered test). Having this much granularity is nice because we can take insights from the individual level and extrapolate them to the group level, whereas it can be hard to go in the other direction in some cases.

However, it would be nice to have features about their families' health histories so we can incorporate the potential effects of genetics on the incidence of heart disease. For example, having information such as "family has a history of heart disease," "immediate family member has had a heart attack," etc. would be very useful. This could help reduce potential confounding in the dataset if having a history of heart disease affects the incidence of heart disease (which domain knowledge dictates it does), and also affects lifestyle behaviors in the dataset, such as physical activity status, smoker status, etc. This would let us use other methods to assess for causality (we had to use matching/IPS due to the likely presence of confounders).

2 Research Question

For this dataset, we are specifically interested in two research questions.

- Are there any underlying relationships between variables? When split into two groups according to another potentially relevant categorical variable from the dataset, are there any statistically significant differences in the means of any independent/controlled variables?
- What is the causal relationship between sleep duration and heart disease?

Multiple hypothesis testing helps to answer the first question because we look for the relationships between one independent variable and multiple other variables. Multiple hypothesis testing allows us to adjust the false discovery rate encountered in conducting multiple single hypothesis testing. By answering this research question, scholars can identify underlying relationships which can be investigated further for medical importance.

Causal inference is the most appropriate method to answer the second research question because it controls the confounding variables, resulting in the most accurate causality between sleep time and heart disease. Answering this question, scholars may use the causality to develop new prevention or treatment for heart disease. For example, if the causality is positive, doctors may add medicines that give people better sleep in heart disease treatment. Another real-life decision this research question helps to make is greater investment in relevant products. For example, if the causality is positive, companies may spend more time on sleeping product research and development to increase sleeping quality and thus prevent heart disease.

3 EDA

For the first research question, since we are interested in relationships between variables, we first visualize the correlation coefficient between all numerical variables using the heatmap. We first convert all binary categorical variables(yes vs. no and male vs. female) into numerical values 1 and 0. Then, using the package seaborn, we drew out the heatmap, which could help us find potential variables that could be confounding or instrumental. Reading the heatmap, we infer that most of the variables are uncorrelated from each other. This may be caused by converting binary categorical variables into 1s and 0s. However, these low correlation coefficients do not indicate 0 causation, we still need to run multiple hypothesis testing and causal inference to verify two variables are indeed uncorrelated or independent of each other. In other words, this heatmap motivates and suggests the answer to the hypothesis testing step.

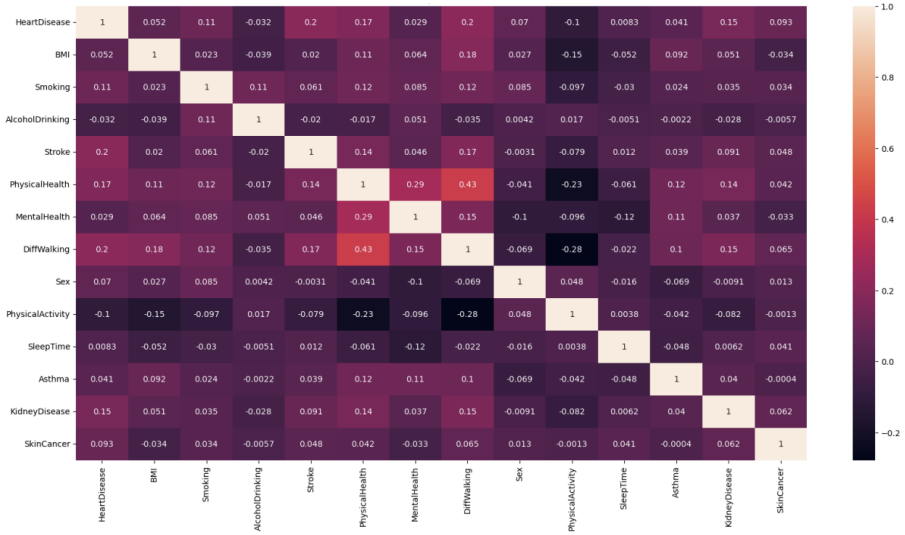

Figure 1: Heatmap for all variables

The 0.29 correlation indicated by the heatmap between physical and mental health made us interested in their relationship. Thus, we visualized this histogram and explored the distributions of Mental Health (numerical) for those physically active versus those not physically active (the split is created via a categorical variable in the dataset). Since this data came pre-cleaned, the only cleaning step we conducted was to turn the Physically Active variable into an integer, either 1 or 0, to reflect the binary results of the question. We notice that, while the results are pretty similar regardless of physical activity, there does appear to be a slight negative correlation between physical activity and mental health. However, this trend does not indicate any causality. We also noticed that a large proportion of both groups said that their mental health was at or below 5 out of 30 (roughly 30% of people). Since this dataset came from phone-call responses, this collection process raised our concern about the volunteer bias - the external validity may be skewed if people who respond to these polls are systematically different than those who do not.

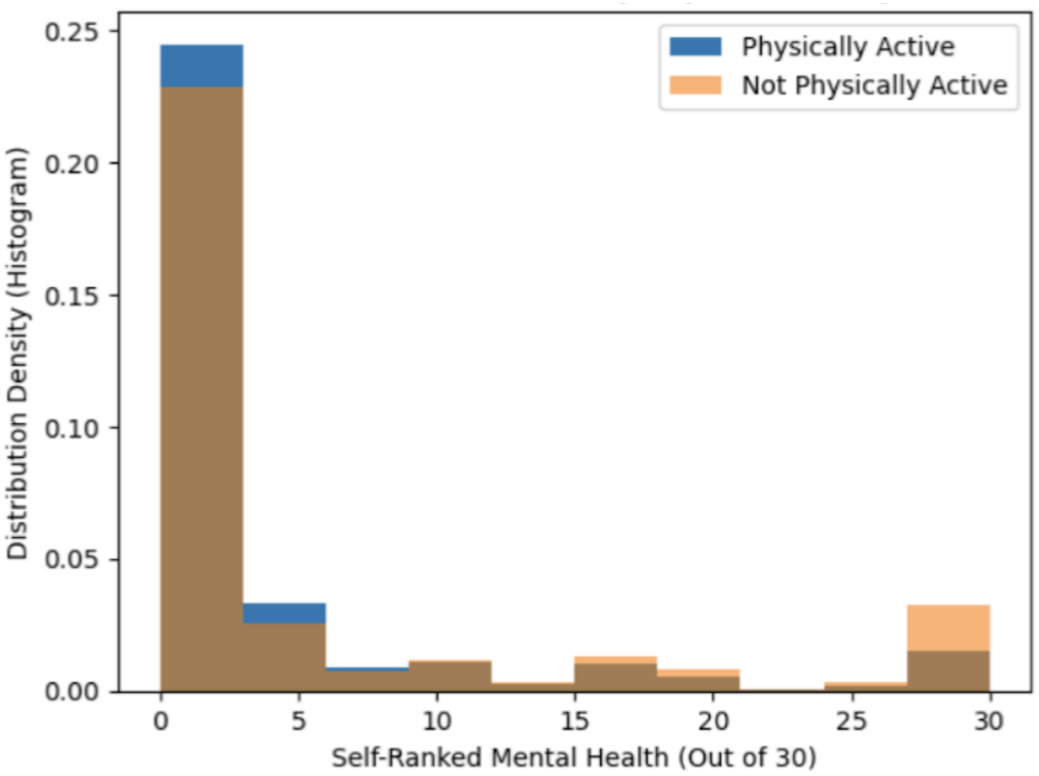

Figure 2: Distribution of Mental Health By Physical Activity Status

For the second research question, we are interested in looking for potential confounders in the causal relationship between sleep time and heart disease. The visualization below shows the distribution of sleep time given one's alcohol-drinking status. According to the visualization, there seems to be a slight difference, between the distribution of sleep time of frequent drinkers and non-frequent drinkers, with both groups having their mean around the 8-9 hours range. Differences such as this motivated us to conduct further research for credible resources online to determine whether certain variables may be considered a potential confounder for the causal inference part of the research.

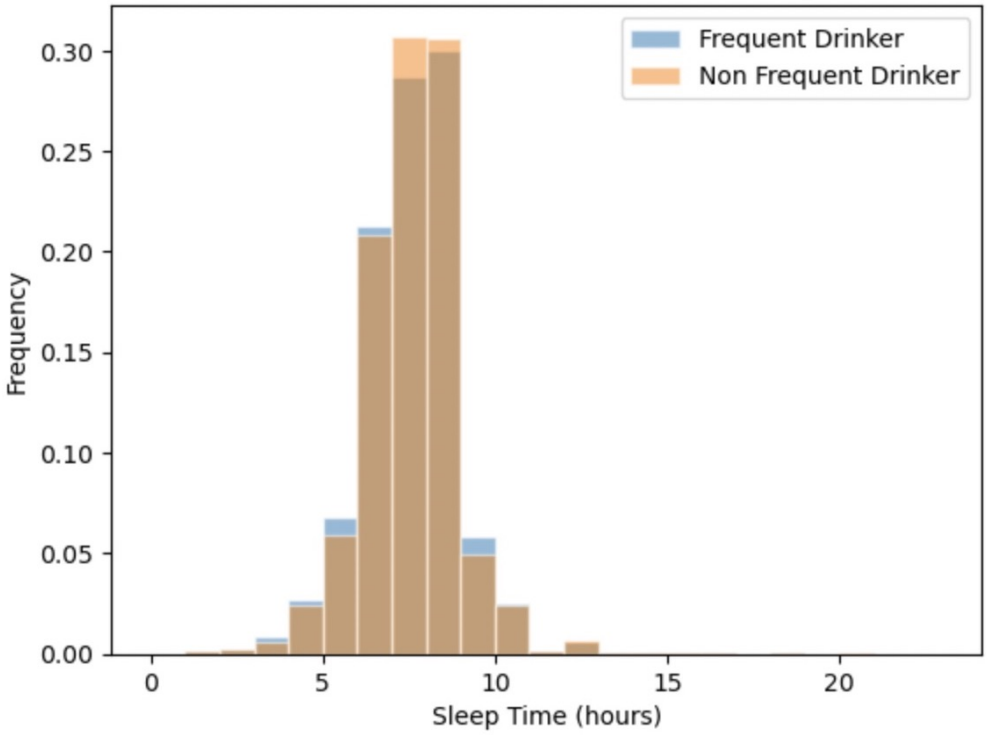

Figure 3: Sleep time distribution for frequent and non-frequent smokers

For numerical variables such as MentalHealth, a histogram is not a good visualization for its relationship with SleepTime, which is also a numerical variable. Therefore, we choose a scatterplot to demonstrate the relationship between two numerical variables. To properly observe trends and eliminate duplicate values, we measured the average mental health value for each possible sleep time value. There seems to be a positive linear association between average mental health and sleep time, with a few outliers (mostly good sleepers with bad mental health) scattered throughout. For cleaning/data editing, we averaged the mental health values for each possible sleep time value. This hides the variance, distribution, and count of values for each sleeping level, but allows us to evaluate for a trend. This will not significantly affect our model/inferences for our research question (due to the nature of the relevance of this visualization described below). Since we want to know the impact of sleep and mental health on heart disease, understanding the relationship between sleep and mental health is also important. However, the positive association does not show causation or correlation. This motivates further research by showing that to determine the individual effects of sleep and mental health on heart disease, we should take steps to carefully control for the other variable since it is likely a confounding variable.

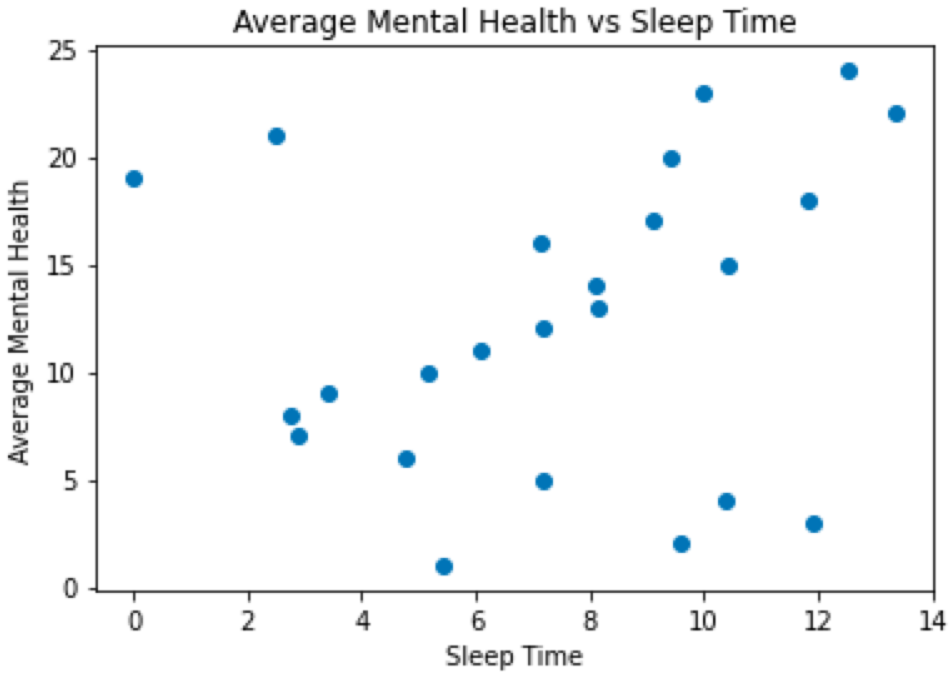

Figure 4: Average mental health condition for different sleep duration

4 Multiple Hypothesis Testing

4.1 Methods

Next, we will be running two-tailed hypothesis tests to determine if there is evidence of underlying relationships within the seven pairs of independent and controlled variables. The null hypothesis for our test will be that there is no difference in the means of a (quantitative) variable when split into two groups according to our other (categorical) variable. (e.g. Is there a difference in mean hours slept for those who drink alcohol and those who do not? Null: there is no difference. Alternate: there is a difference). It makes sense for us to test multiple hypotheses because we want to know if there are any variables (out of the variables that are not our independent/dependent variables) that have underlying relationships with our independent variables. Running one hypothesis test would only give us information about one pair of variables. We used 400 permutations of the dataset to run our hypothesis tests (using more caused our kernel to crash). We are testing each hypothesis using an A/B test because we are dealing with two variables at a time and trying to assess whether there is any statistical difference in one of them while it is being controlled by the other (eg. sleep time for those who drink and those who do not) We will be using Bonferroni and the BH procedure to control for FWER and FDR (respectively). These error rates follow their standard definitions and we are controlling them to ensure the validity of our results as we test multiple hypotheses (and want to avoid T1 errors for individual tests and the grouping of tests). This is described more in subsequent sections.

4.2 Result

Upon running two-tailed hypothesis tests for seven pairs of independent and controlled variables, we obtained p-values that were all less than our significance level (0.05), leading us to reject our null hypothesis and conclude that there is a statistically significant difference between our quantitative variables when split up according to their respective categorical variables. We used the Bonferroni correction to control the family-wise error rate and the Benjamini-Hochberg correction to control false discovery rate (these follow their standard statistical definitions described in class and their implications described above). As a refresher,

FWER is the probability of making 1 or more T1 errors (false discoveries) among our multiple hypothesis tests, so controlling this controls the probability of making a T1 error. FDR is the rate at which null features are falsely assessed as significant, so controlling this also controls T1 errors but in a different way - it caps the FDR at the decided alpha (we used .05). According to the results from our corrections, our initial results were validated except for the tests involving sleep time versus physical activity and sleep time versus alcohol drinking.

4.3 Discussion

As just stated, our initial results were validated except for the tests involving sleep time versus physical activity and sleep time versus alcohol drinking after applying the corrections. The results of our correction led to a failure to reject the null hypothesis for these two pairs, as our p-values were not lower than the updated cutoff. The results for the other five pairs remained relevant since they were validated by all three procedures. Decisions that stem from the relationships between the individual variables include suggesting a further investigation into those relationships for potential medical significance, as well as any other real-life decisions that need to be made related to the relationships. As a group, we can assess that there are significant relationships between our variables - meaning that our decision to heavily investigate confounders is extra necessary (we would have done this anyways). Our results are limited in that they only let us know the differences between means and if there is a statistically significant difference there, whereas there could be very practically significant differences in the underlying distributions that would require different tests to discover. We avoided p-hacking by not testing the same hypothesis in multiple ways and by leveraging the BH and Bonferroni corrections. If we had more data, we would test the relationships between more potentially relevant variables. We could also attempt to determine causality between some of them (which is what we did for our other research question).

	p_vals	test	bon_decisions	bh_decisions	naive_decisions
0	0.0000	Mental Health and High Sleep	1	1	1
1	0.0075	Sleep Time and Alcohol Drinking	0	0	1
2	0.0000	Mental Health and Alcohol Drinking	1	1	1
3	0.0325	Sleep Time and Physical Activity	0	0	1
4	0.0000	Mental Health and Physical Activity	1	1	1
5	0.0000	Mental Health and Asthma	1	1	1
6	0.0000	Sleep Time and Asthma	1	1	1

Figure 5: Adjusted P-values

5 Causal Inference

For the second research question, we decided first to use both instrumental variable regression and matching separately to find the causal relationship between sleep and heart disease. Then, we cross-checked these two results to form a stronger conclusion.

5.1 Methods

Based on the research question, the treatment for this study is the amount of sleep each patient self-reported, and the outcome is the binary result of whether this patient has heart disease.

- **SleepTime** (discrete numerical variable): On average, how many hours of sleep does this patient get in a 24-hour period?

- **HeartDisease** (binary categorical variable): Yes for respondents who reported having coronary heart disease or myocardial infarction.

Conducted information $research^{1-9}$ on the relationship between all variables and heart disease, we first identified 0 collider and 6 confounding variables that impact both sleep and heart disease. They are BMI, smoking, alcohol, age, diabetes, and physical activity.

- **BMI** (continuous numerical variable): Body mass index
- **Smoking** (binary categorical variable): Has this patient smoked at least 100 cigarettes in his/her entire life?
- **AlcoholDrinking** (binary categorical variable): Yes for adult men having more than 14 drinks per week and adult women having more than 7 drinks per week
- **AgeCategory** (discrete categorical variable): Fourteen-level age categories, leveled as 18-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75-79, 80 or older
- **Diabetic** (binary categorical variable): Whether this patient been told have diabetes?
- **PhysicalActivity** (binary categorical variable): Yes for adults who reported doing physical activity or exercise during the past 30 days other than their regular job

After identifying these 6 confounders, we proceeded by utilizing 2 different approaches to assume unconfoundedness.

5.1.1 Matching

First, we have decided to use matching to research the causal relationship between sleep time and heart disease. Through preliminary research, we have determined that there are 6 confounding variables that would impact this relationship. We first checked the number of units in the entire dataset with at least one matching case (have similar confounding conditions with at least one other unit), and obtained that 99.8% of the units in the dataset have a matching case. This means that the majority of the units will be considered in the matching process. We henceforth decided matching a viable method to proceed.

```
trim_test = trim_data[["BMI", "Smoking", "AlcoholDrinking", "AgeCategory", "Diabetic", "Asthma"]]
trim_test.duplicated(keep=False).sum() / len(trim_data)
```

```
0.9984771494238497
```

Figure 6: Percentage of Units with Matching Confounders

Proceeding from here, we decide to take the groups of people with similar confounders that have the number of units above the 10th percentile based on the distribution of the number of units within the group. This eliminates the groups with too few people having similar confounders to make the final results more conclusive.

In order to divide the dataset into two categories – those who were untreated and those who were treated – for matching, we have converted SleepTime, originally a numerical variable, into a categorical variable. We achieved such conversion through the official guidelines of the appropriate sleep time given age groups provided by the CDC. For example, the CDC recommended that the appropriate sleep time for the age group 18 - 24 to be at least 7 hours. We would therefore mark those who are in the 18 - 24 age group and sleep less than 7 hours (don't have appropriate sleep time) to be "treated" (1), and those who sleep more than 7 hours within the age group to be "untreated" (0). The reason that we chose those whose sleep time does not match the amount recommended by CDC to be the treated group and hence marked as "1" is because, by common sense, we deem those who have worse sleep quality more likely to get heart disease. Thus, the ATE will be positive if people who have worse sleeping times with the same confounding conditions have more heart disease cases among them. We constructed the following function to select units from those who were treated and those who weren't with similar confounding conditions.

We have also converted the BMI data given in the dataset from a numerical variable into a categorical variable. This guarantees that there won't be too many units in the dataset without a match of their exact BMI, making the results of matching with too much inaccuracy. We have converted the BMI data with the following criteria: Category 1: $BMI \leq 18.5$; Category 2: $18.5 \leq BMI \leq 24.9$; Category 3: $25 \leq BMI \leq 29.9$; Category 4: $BMI \leq 30$

A snippet of the resulting dataframe with properly engineered features and converted treatment and untreated groups looks like the following:

	HeartDisease	SleepTime	BMI	Smoking	AlcoholDrinking	AgeCategory	Diabetic	Asthma
0	0	5.0	1.0	1	0	55-59	1	1
1	0	7.0	2.0	0	0	80 or older	0	0
2	0	8.0	3.0	1	0	65-69	1	1
3	0	6.0	2.0	0	0	75-79	0	0
4	0	8.0	2.0	0	0	40-44	0	0
...	...	...	...	...	...	...	...	...
319790	1	6.0	3.0	1	0	60-64	1	1
319791	0	5.0	3.0	1	0	35-39	0	1
319792	0	6.0	2.0	0	0	45-49	0	0
319793	0	12.0	4.0	0	0	25-29	0	0
319794	0	8.0	4.0	0	0	80 or older	0	0

319795 rows × 8 columns

Figure 7: Processed Data Snippet

We then constructed a function to calculate the average treatment effect (ATE) for the treatment and untreated groups with these confounders, and performed calculations of ATE for all the possible permutations of conditions for treatment and untreated groups with the following procedures. An example calculation can be seen as the following:

Treatment group with a set permutation of confounders was selected.

```
treat_test = return_treatment(bmi = 1, smoking = 0, alcohol = 0, diabetic = 1, asthma = 0, age = '80 or older')
treat_test
```

	HeartDisease	SleepTime	BMI	Smoking	AlcoholDrinking	AgeCategory	Diabetic	Asthma
966	0	1.0	1.0	0	0	80 or older	1	0
8409	1	1.0	1.0	0	0	80 or older	1	0
16316	0	1.0	1.0	0	0	80 or older	1	0
18222	1	1.0	1.0	0	0	80 or older	1	0
19803	0	1.0	1.0	0	0	80 or older	1	0
45745	0	1.0	1.0	0	0	80 or older	1	0
62583	0	1.0	1.0	0	0	80 or older	1	0
108008	0	1.0	1.0	0	0	80 or older	1	0
136368	0	1.0	1.0	0	0	80 or older	1	0
149960	0	1.0	1.0	0	0	80 or older	1	0
221890	0	1.0	1.0	0	0	80 or older	1	0
261572	0	1.0	1.0	0	0	80 or older	1	0
269110	0	1.0	1.0	0	0	80 or older	1	0
285417	0	1.0	1.0	0	0	80 or older	1	0
311424	0	1.0	1.0	0	0	80 or older	1	0

Figure 8: Treated Units with Certain Confounders

Untreated group with a set permutation of confounders that matches the treatment group above was selected.

```
untreated_test = return_untreated(bmi = 1, smoking = 0, alcohol = 0, diabetic = 1, asthma = 0, age = '80 or older')
untreated_test
```

	HeartDisease	SleepTime	BMI	Smoking	AlcoholDrinking	AgeCategory	Diabetic	Asthma
15513	0	0.0	1.0	0	0	80 or older	1	0
52667	0	0.0	1.0	0	0	80 or older	1	0
54623	1	0.0	1.0	0	0	80 or older	1	0
55975	1	0.0	1.0	0	0	80 or older	1	0
62195	0	0.0	1.0	0	0	80 or older	1	0
83233	0	0.0	1.0	0	0	80 or older	1	0
91407	0	0.0	1.0	0	0	80 or older	1	0
126055	1	0.0	1.0	0	0	80 or older	1	0
152368	0	0.0	1.0	0	0	80 or older	1	0
168413	1	0.0	1.0	0	0	80 or older	1	0
220366	1	0.0	1.0	0	0	80 or older	1	0
221753	0	0.0	1.0	0	0	80 or older	1	0
279837	0	0.0	1.0	0	0	80 or older	1	0

Figure 9: Untreated Units with Matching Confounders

Average Treatment Effect for the two groups was calculated.

```
calculate_ate(treat_test, untreated_test)
```

```
-0.2512820512820513
```

Figure 10: Sample ATE Calculation

Finally, we have decided to eliminate those groups with overly big confounding groups and overly small groups with similar confounders so as to not overwhelm the matching process. We have plotted the distribution of

the number of people for all possible permutations of confounders. The distribution of the treatment group is as the following:

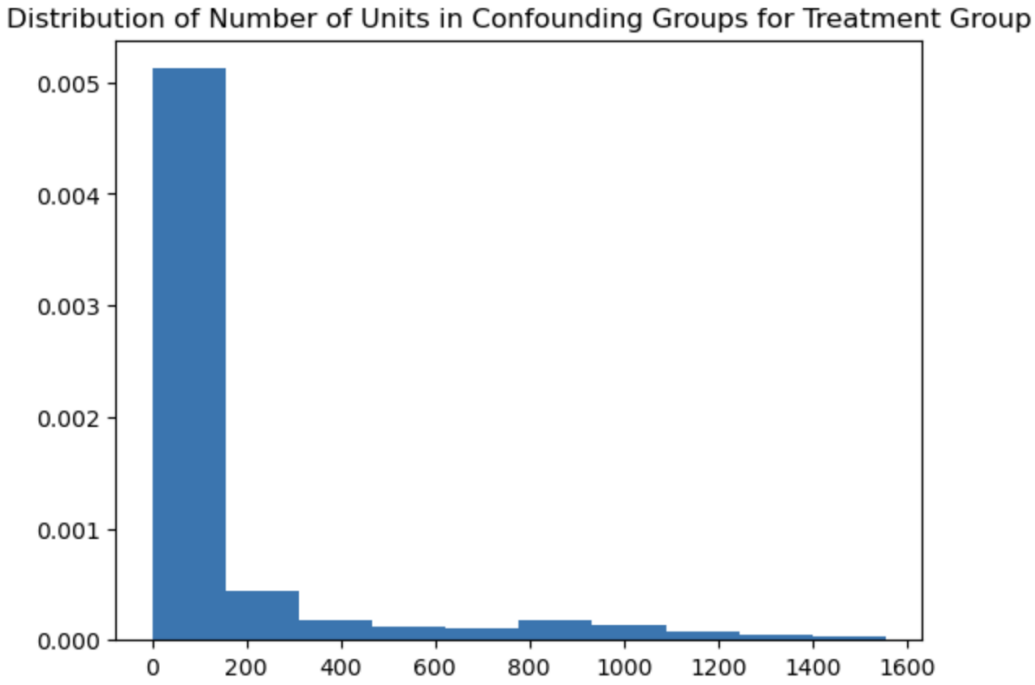

Figure 11: Distribution of Number of Units in Confounding Groups for Treated Group

The distribution of the numbers of people in an untreated group is as the following:

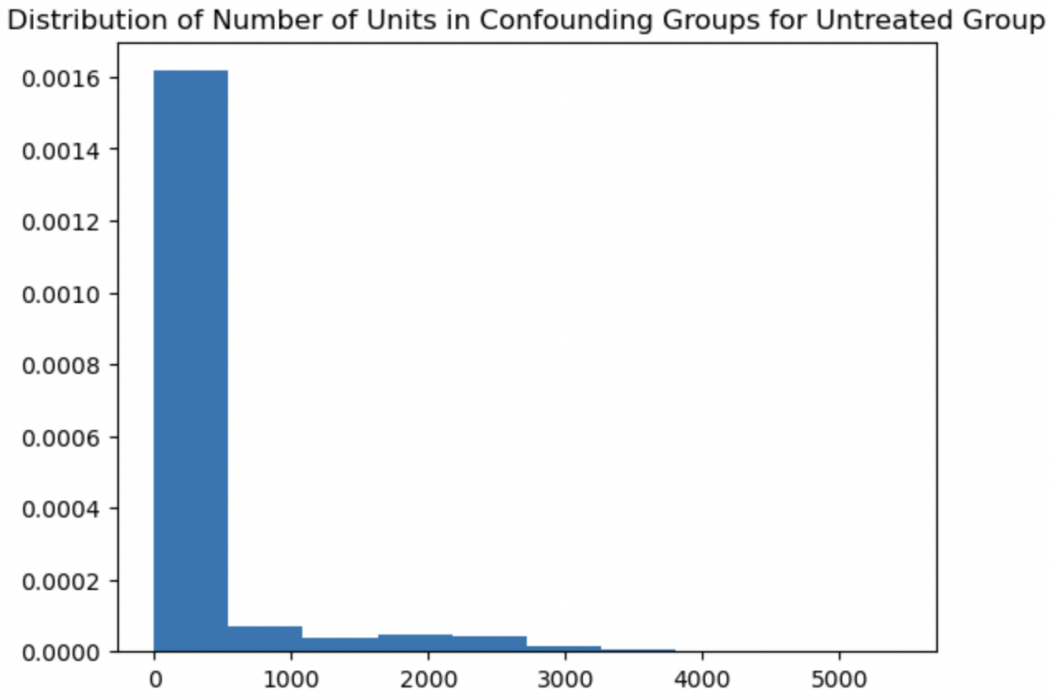

Figure 12: Distribution of Number of Units in Confounding Groups for Untreated Group

Based on the distribution above, a final condition we have decided to use to select the data is that we have decided to take those treated and untreated confounding groups only within the 10th to 90th percentile of their respective distributions above, so as to avoid overly big and overly small groups.

5.1.2 Regression with Instrumental Variable

After identifying 6 confounding variables, we researched the remaining variables' relationships with the treatment variable SleepTime and their relationships with these 6 confounders. This additional $research^{10-15}$ aims to find variables that are independent of all 6 confounders but impact the treatment. Luckily, based on the research result, patients' self-reported physical health condition can be used instrumental variable because it satisfies both requirements – that it is independent of all the remaining variables, except that it affects heart health through sleep time and only through sleep time. That is, We adjust the confounders by finding the instrumental variable and using it to conduct a 2SLS regression.

- **PhysicalHealth** (discrete numerical variable): The number of days in the past 30 days when this patient suffers physical illness and/or injury.

To conduct the 2SLS regression, we first fit SleepTime from PhysicalHealth to get the prediction $\hat{SleepTime}$. Then, we fit the outcome HeartDisease from the prediction $\hat{SleepTime}$ to obtain the causal relationship coefficient $\hat{\tau}$.

5.2 Result

Despite two different approaches, both matching and instrumental variable regression result in similar conclusions: enough sleep decreases the chance of having heart disease. In other words, lack of sleep increases the chance of having heart disease. Besides the ATE for all patients, we also observe the ATEs across different age categories. From these calculations, we conclude that the older the group, the stronger this causality.

5.2.1 Matching

To add additional value and for this research to have significance to our everyday life, we decided to plot the resulting ATEs based on different age groups. We will then be able to see whether aging has an impact on the causal relationship between sleep time and heart disease. We eventually used a box plot to visualize such distribution, and the results were presented below:

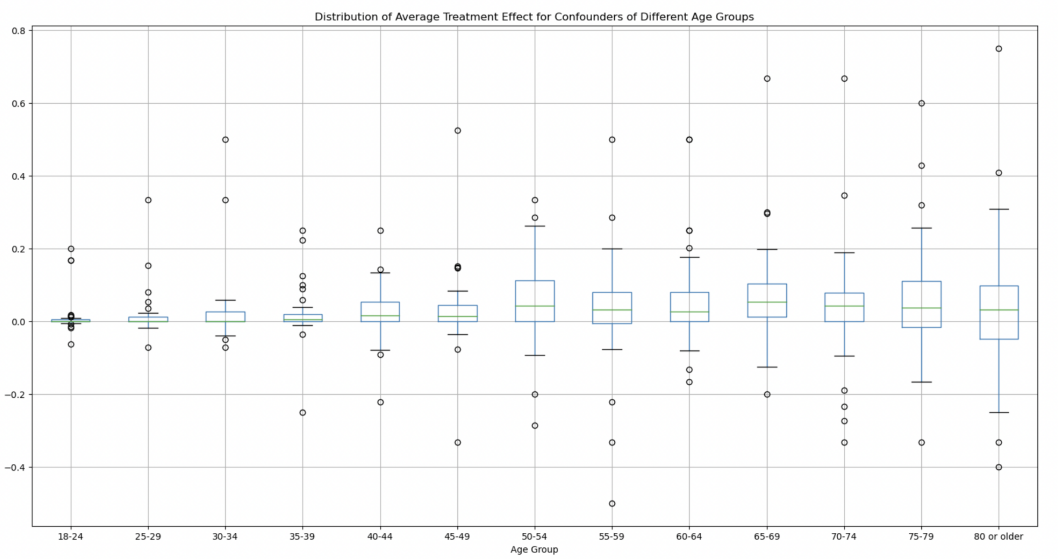

Figure 13: Distribution of Average Treatment Effect for Confounders of Different Age Groups

From the above figure, it can be seen that there was a general increase in the mean for the distribution of each age group. The spread of the distribution is also increasing, and there is a general increase in the maximum value of ATE attained as the age group increases. That being said, as the spread increase, there seems to be a trend in the increase in negative ATE as well. This shows that as age increase, the likelihood of getting heart disease increase regardless of one's sleep time. However, this does not nullify the causal relationship, as there is a greater increase in positive ATE compared to negative ones.

5.2.2 Regression with Instrumental Variable

For 2SLS regression, the coefficient estimated $\hat{\tau}$ is -0.5418 with t-statistic = -97.982, p-value = 0, and confidence internal doesn't include 0(-0.553, -0.531). These indices indicate the statistical significance of the result. We interpret this result as one hour more of sleep leads to a 54.18% decrease in having heart disease.

OLS Regression Results

Dep. Variable:	HeartDisease	R-squared:	0.029
Model:	OLS	Adj. R-squared:	0.029
Method:	Least Squares	F-statistic:	9600.
Date:	Tue, 06 Dec 2022	Prob (F-statistic):	0.00
Time:	12:29:50	Log-Likelihood:	-41684.
No. Observations:	319795	AIC:	8.337e+04
Df Residuals:	319793	BIC:	8.339e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.9309	0.039	100.155	0.000	3.854	4.008
PhysicalHealth	-0.5418	0.006	-97.982	0.000	-0.553	-0.531

Omnibus:	179686.010	Durbin-Watson:	1.974
Prob(All): 0.000	Jarque-Bera (JB):	1007210.053	
Skew:	2.852	Prob(JB):	0.00
Kurtosis:	9.562	Cond. No.	583.

Figure 14: 2SLS regression result table

Then, we take a look at the causal impact of hours of sleep on heart disease across different age groups. The results conducted following the same 2SLS regression procedures are listed below.

Age Group	Coeffecients
18-24	-0.053572
25-29	-0.045253
30-34	-0.061575
35-39	-0.089865
40-44	-0.170802
45-49	-0.197462
50-54	-0.258853
55-59	-0.330907
60-64	-0.404624
65-69	-0.463205
70-74	-0.755524
75-79	-61.090220
80 or older	-7.295324

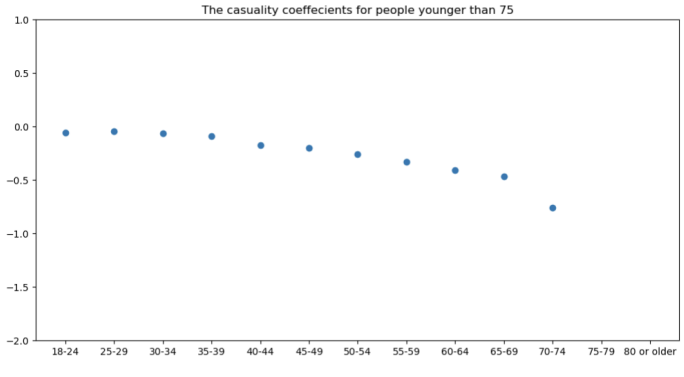

Figure 15: The coefficients table along with the visualization

The chart above shows that the coefficients decrease as the age group increases. That is, the causality of sleep to heart disease increases as people get older. More specifically, the older the patient, the greater sleep impacts heart disease prevention.

The decrease of causality from 70-74 to 80 or older indicates some uncertainty within the estimation. However, this uncertainty is not evidenced against our hypothesis that a positive causal relationship exists between sleep and heart disease. After all, these two coefficients are still negative, indicating that more sleep causes the chance of getting heart disease to decrease.

5.3 Discussion

The first limitation of the matching method comes from the fact that the Average Treatment Effect may never be observed, and hence the matching method can only be considered an estimation of the real ATE.

Another limitation of the matching method may come from the dataset being simplified in order to utilize as many units during the matching process as possible. For example, we have recategorized features such as sleep time and BMI from numerical to categorical variables. Such categorization may also have an impact on the final results as different categorizations. Another limitation for the matching process comes from some data being omitted during the process. There are two sources for such omissions. The first one is that we have purposefully omitted the confounding groups with overly big and overly small units inside them, based on the distribution of the number of people in each group. Another source of omission comes from certain individuals in the treatment/untreated group do not have a matching unit for the same set of confounders, and their cases are excluded from the matching process as a result. We are confident with the result of our causal inference procedure, as both matching and instrumental variables yielded similar trends and results. However, another limitation comes from the fact that the data we are working on is a processed and simplified version of the CDC data, where only indicators for major health conditions were included in the version we were working on. An assumption that was made in the process of using the simplified version of the data is that we assumed the confounders that we eventually used to perform causal inference analysis were the only confounders that matter in this causal relationship. That is to say, we need more features and feature engineering, as well as relevant domain knowledge to be able to determine important confounders and yield a more accurate result.

6 Conclusion

Overall, our hypothesis tests showed that the numerical variables have different means when they are split into groups according to other Boolean health-data variables (at least for the 7 relationships we tested). However, the application of BH and Bonferroni corrections made 2 of these 7 tests not significant at the 95% confidence level when controlling for their associates' error rates (BH controls FDR and Bonferroni controls FWER). A key takeaway from our research is that there are many different underlying relationships between the health metrics in our dataset. While the hypothesis tests did not prove a causal relationship, the presence of these relationships makes us extra cautious and deliberate to account for confounders in determining causation from health factors on heart disease. This makes sense since we intuitively hypothesized that these four factors are important enough that a significant change in one will have a noticeable effect on others. Due to the sampling accuracy and the broad nature of the study (factors like sleep time and mental health are universally applicable), our results should be generalizable to larger populations.

For the causal inference section, we did both matching and instrumental variable 2SLS regression to find out the causal relationship between SleepTime and HeartDisease. Our key findings are:

- After adjusting all confounders, sleep has a negative causal relationship with heart disease. The more you sleep, the less likely you get heart disease.
- This causal relationship is weaker for young people and stronger for older people.

Based on the above conclusion, pharmaceutical companies may spend more time and money on developing medicine that helps people get enough sleep. Additionally, since this causality is stronger for older people, these companies may pay extra attention to developing medicine specifically suitable for the elderly to ensure their sleeping quality and thus prevent heart disease.

While the results of identifying the existence of underlying relationships using multiple hypothesis testing may not have as significant of a call to action as that of the causal inference section, we can still suggest some follow-up actions. Given that we are finding relationships between variables, investigating causal effects would help with a general understanding of medical science and medical treatment. (E.g. if sleep time and alcohol drinking have a causal relationship, and sleep time is important for other health outcomes, knowledge from investigating this relationship could improve medical advice and health outcomes.) Also, given some of the relationships discovered, investigating for additional relationships between health behaviors and outcomes must be completed to increase medical knowledge for the same reasoning just described.

All of our research was centered on the sole CDC survey and we did not merge any other sources. This prevented any further EDA complications since there were not varying sampling procedures, survey questions,

time periods, and other factors that we needed to control for, and bolstered the generalizability of our results. One consequence of not using other data sources could be the limited scope of the sample. Although 400,000 respondents is a significant amount, the US has over 200,000,000 adults, meaning we studied less than 0.2 percent of our overall population. This necessitates a key assumption: that our dataset represents the actual population. If this is not true, the results could be skewed. While we cannot fully validate this, we did our best by checking estimates for population parameters (e.g. the US gender split) from the data for the data that is possible to do this for, and we found it to be pretty close the real population parameters (indicating that it likely is a representative sample, although we must keep this assumption in mind regardless).

A limitation in our data was the number of permutations we could run given our available computing power. We ran our hypothesis tests using 400 permutations of the data when the possible number of permutations was several orders of magnitude larger. This is a very key assumption in this part of the project: that the 400 permutations is sufficient in simulating the data. Another limitation was the current status of domain knowledge on heart disease. Since we used domain knowledge from medical research (see citations and other sections of this paper for more details) to pick confounders, we could have missed confounders that are fully unrelated to everything in the data. This assumption is paramount in the validity of our results (and is why we researched it to determine the best possible list of variables to include as potential confounding variables).

Examining the 2SLS regression from the casual inference section, we also acknowledged its limitation of using least square regression for a binary outcome. That is, the least-square regression model we designed may not be the most appropriate model because it fails to bound the result between 0 and 1. Instead, we should try using the instrumental variable in logistic regression.

All these limitations render space for further work. To address the lack of generality issue, scholars could do similar phone-call surveys in other countries to generalize conclusions. Also, scholars who are interested in the causal relationship between sleep quality, or sleep duration, and heart disease will now consider sleep duration as the confounder. This careful examination of confounding variables will produce a more accurate research result.

Word Count: 4965

Reference Page

[1] Akil, Luma, and H Anwar Ahmad. “Relationships between obesity and cardiovascular diseases in four southern states and Colorado.” *Journal of health care for the poor and underserved* vol. 22,4 Suppl (2011): 61-72. doi:10.1353/hpu.2011.0166

[2] Do, Young Kyung. “Causal Effect of Sleep Duration on Body Weight in Adolescents: A Population-based Study Using a Natural Experiment.” *Epidemiology (Cambridge, Mass.)* vol. 30,6 (2019): 876-884. doi:10.1097/EDE.0000000000001086

[3] Leary, Peter J. “Causality, Correlation, and Cardiac Disease: Does Smoking Cause Cardiac Hypertrophy and Diastolic Dysfunction?.” *Circulation. Cardiovascular imaging* vol. 9,9 (2016): e005441. doi:10.1161/CIRCIMAGING.116.005441

[4] Palmer CD, Harrison GA, Hiorns RW. Association between smoking and drinking and sleep duration. Ann Hum Biol. 1980 Mar-Apr;7(2):103-7. doi: 10.1080/03014468000004111. PMID: 7425536.

[5] Piano, Mariann R. “Alcohol's Effects on the Cardiovascular System.” *Alcohol research : current reviews* vol. 38,2 (2017): 219-241.

[6] Stein, Michael D, and Peter D Friedmann. “Disturbed sleep and its relationship to alcohol use.” *Substance abuse* vol. 26,1 (2005): 1-13. doi:10.1300/j465v26n01_01

[7] Britton, A., Fat, L.N. & Neligan, A. The association between alcohol consumption and sleep disorders among older people in the general population. *Sci Rep* 10, 5275 (2020). https://doi.org/10.1038/s41598-020-62227-0

[8] Cukic, Vesna et al. “Sleep disorders in patients with bronchial asthma.” *Materia socio-medica* vol. 23,4 (2011): 235-7. doi:10.5455/msm.2011.23.2

[9] Xu, Mingzhu et al. “Asthma and risk of cardiovascular disease or all-cause mortality: a meta-analysis.” *Annals of Saudi medicine* vol. 37,2 (2017): 99-105. doi:10.5144/0256-4947.2017.99

[10] Patyar S, Patyar RR. Correlation between Sleep Duration and Risk of Stroke. J Stroke Cerebrovasc Dis. 2015 May;24(5):905-11. doi: 10.1016/j.jstrokecerebrovasdis.2014.12.038. Epub 2015 Mar 25. PMID: 25817615.

[11] Shiozawa, Masahiro et al. “Association of Body Mass Index with Ischemic and Hemorrhagic Stroke.” *Nutrients* vol. 13,7 2343. 9 Jul. 2021, doi:10.3390/nu13072343

[12] Burgard, Sarah A, and Jennifer A Ailshire. “Gender and Time for Sleep among U.S. Adults.” *American sociological review* vol. 78,1 (2013): 51-69. doi:10.1177/0003122412472048

[13] Chinwong, Dujrudee et al. “A Comparison of Gender Differences in Smoking Behaviors, Intention to Quit, and Nicotine Dependence among Thai University Students.” *Journal of addiction* vol. 2018 8081670. 24 Oct. 2018, doi:10.1155/2018/8081670

[14] Yacoub, Rabi et al. “Association between smoking and chronic kidney disease: a case control study.” *BMC public health* vol. 10 731. 25 Nov. 2010, doi:10.1186/1471-2458-10-731

[15] Arafa A, Mostafa A, Navarini AA, Dong JY. The association between smoking and risk of skin cancer: a meta-analysis of cohort studies. Cancer Causes Control. 2020 Aug;31(8):787-794. doi: 10.1007/s10552-020-01319-8. Epub 2020 May 27. PMID: 32458137.