# Data 102 Final project report

By Alvin Han, Ian Howe, Jun Kee Kim Kim, Cary Kuang

# Abstract

The Center for Disease Control and Prevention (CDC) maintains the U.S. Chronic Disease Indicators dataset, containing state-specific data of chronic illness prevalence as well as relevant policies and regulations. The CDC also publishes daily air quality data through the National Environmental Public Health Tracking Network to monitor environmental exposures, which was used in conjunction with the Chronic Disease Indicators to test the significance of PM2.5 concentration on six health concerns: mental health, chronic kidney disease, lung cancer, chronic obstructive pulmonary disease, cardiovascular disease, and asthma. When controlling for the FWER and FDR with the Bonferroni correction and the Benjamini-Hochberg process, COPD was found to be significant for both methods while chronic kidney disease was also significant under only the Benjamini-Hochberg process. In addition to these hypothesis tests, we wanted to establish whether there was a causal relationship between smoking tobacco and COPD, one of the diseases that was significantly related to air quality. Due to confounding variables that could affect both the treatment and the outcome, we used whether or not a state had strong laws regarding the sale of tobacco products as an instrument for tobacco consumption. Despite expectations, we found no causal relationship between tobacco use and COPD with our data. However, there are some shortcomings of our model that could be addressed given more resources for a more accurate model, and the results may differ from ours.


# Background

Air pollution is a major public health concern, with numerous studies demonstrating its harmful effects on human health. One of the most significant components of air pollution is fine

particulate matter, or PM2.5. It is composed of tiny particles that are less than 2.5 micrometers in diameter. These particles are small enough to penetrate deep into the lungs, where they can cause a range of respiratory and cardiovascular problems. A growing body of research has focused on the relationship between PM2.5 concentration and respiratory diseases, such as chronic pulmonary disease (COPD), asthma, lung cancer, cardiovascular disease, kidney disease, and in turn, mental health. WHile the precise mechanisms by which PM2.5 exposure leads to these conditions are not yet fully realized, there is strong evidence to suggest that increased exposure to PM2.5 is associated with a higher risk of respiratory disease among adults.

This study aims to investigate the correlation between PM2.5 concentration and the prevalence of respiratory diseases among adults. Specifically, we will analyze data compiled from CDC. By shedding light on this critical issue, this study has the potential to inform public health policy and promote bipartisan inventions that reduce exposure to PM2.5 and improve respiratory health outcomes.

## Data Overview

For us to better understand the effects that PM2.5 concentrations have on the general population of each state, we chose to use causal inference and multiple hypothesis testing. The data that we conducted our study on was provided by the CDC dataset: Annual State-Level U.S. Chronic Disease Indicators, subsequently filtered for each condition that we were interested in (COPD, Asthma, Cardiovascular disease, Tobacco, Lung Cancer, Mental health) and the CDC dataset: Daily Census-Tract PM2.5 Concentrations.

The first dataset was collected by CDC from local hospitals, laboratories, and doctor offices, and the method of collection was never stated but it was most likely sourced from the

participants' medical records. While it is not clear whether or not the CDC had explicit consent from each participant for their data to be collected and used in this manner, we can reasonably assume that CDC only utilized the data that patients consented for the hospital to have on their medical records. While California mandates that a patient's medical record is private, this law differs in other states, and even then federal bureaus such as the CDC have rights to access this data. Regarding the dataset itself, we took the location of the participant and whether or not this participant has the condition (post-filtering) from each row, and contrasted this information with the second dataset. All information was collected based on the year of the collection.

The second dataset was also collected by the CDC through a daily census on the air quality of each state. The data itself was gathered from the Air Quality System database, and Pollutant Occurrence Codes, and was collected daily from January 1, 2001 to December 31, 2014. The data are daily 24-hour average PM2.5 concentrations calculated on a 12 km x 12 km grid for the continental United States. The scope of the dataset are all the census tracts in the United States, and the dataset was created for professional use to compare air quality and health outcomes. As this data was collected mostly by government agencies and organizations that have federal oversight, the data collected are for public use and no violation of consent would be present as the data is not private. When we used it within our study, the data utilized were the state that the PM2.5 average was calculated for, and the PM2.5 average of the respective state. These were contrasted with the information we gathered previously from the CDC disease dataset to give us a better idea on whether or not the average PM2.5 level of a state has a statistically significant impact on the disease rate of that state.

While the dataset provided by the CDC is acceptable for what we want to learn in this study, there are some limitations to the data that might impact the study and possible

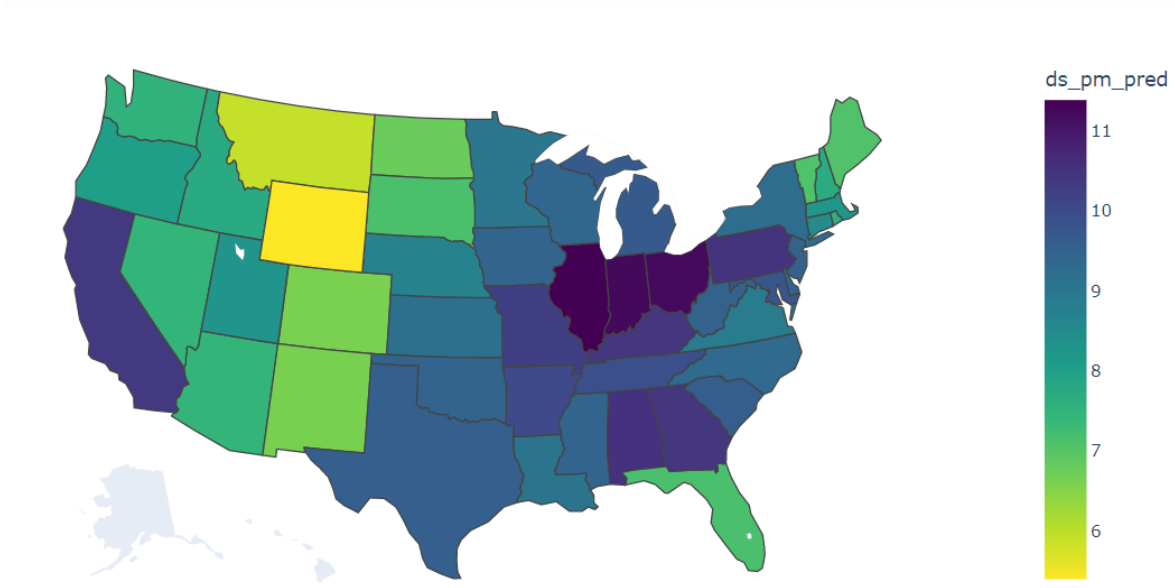
conclusions. The age of the participants in the first dataset are all at least 18 years old. This excludes a very large portion of the population, which is especially concerning when considering the younger population is more susceptible to develop chronic illnesses from poor air quality. Another limitation of the first dataset is that it is susceptible to response and non-response bias. There are patients that are simply not documented in the medical records, whether due to socioeconomic status or other outstanding circumstances. This is crucial as the poor are more likely to live in disadvantaged areas with heavy pollution and more likely to not be able to afford to go to a doctor for conditions developed from said pollution. For participants that do exist within the medical system and thus the survey, there are possibilities of response bias resulting from patients lying to the doctors about their symptoms, either from a fear of treatment itself or the cost of treatment, which results in them appearing as having no diseases or being misdiagnosed.

The listed limitations of the dataset will most likely affect the results of our study, due to our analysis being heavily dependent on the dataset, and our lack of other sources of these data. We will keep the downsides in mind when arriving at our conclusions.

# Exploratory Data Analysis

*PM2.5 concentrations:*

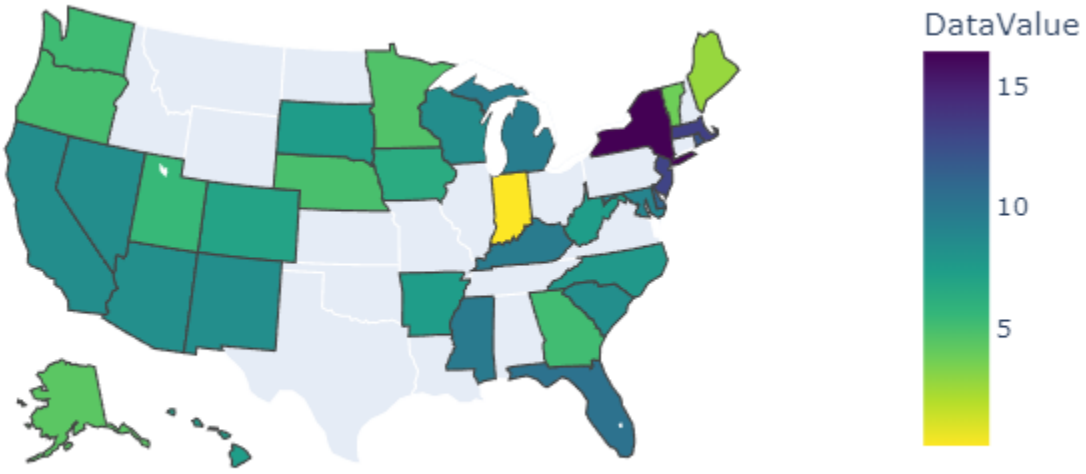
*Average PM2.5 concentrations per 24 hours in μg/m³*




We first mapped PM2.5 concentrations on a choropleth map to see what general trends we could find on air pollution levels across the United States. There is only data for the contiguous United States, and it seems like there are higher levels of PM2.5 concentrations in the midwest with air quality gradually improving as it gets close to the coasts. The one exception to this is California, which has a predicted average PM2.5 concentration of approximately 11μg/m³ compared to the much lower levels of the other pacific states. For our initial exploratory data analysis, we decided to also visualize asthma hospitalizations and lung cancer prevalence to compare to PM2.5 concentration levels. This is because out of all of the health concerns in our analysis, we originally expected these two to have the highest correlation with air quality, and wanted to see if we could observe similar trends on the visualizations.

*Asthma:*

*Average Asthma Hospitalizations per 10,000*



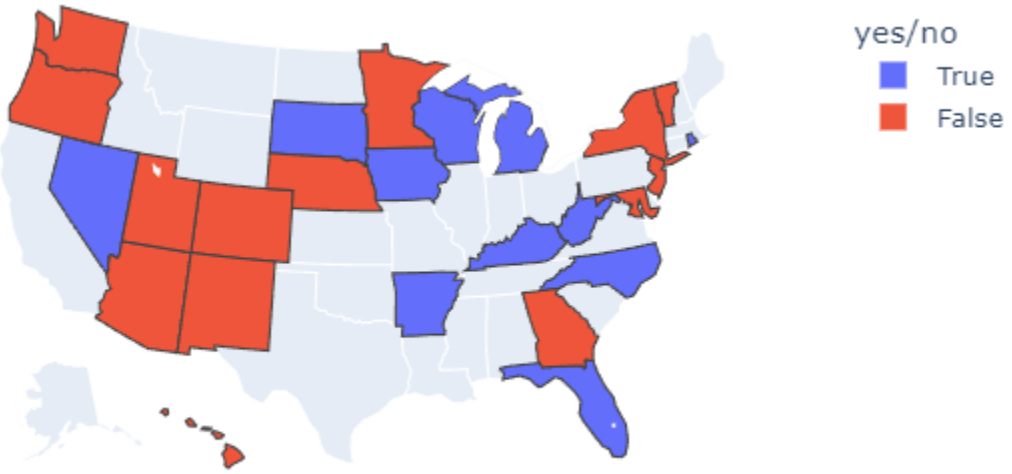

The given choropleth map has values describing the average number of hospitalizations due to asthma per 10,000 people, per state, from 2010 to 2023. The data was first filtered to get rid of NaN values, and then we grouped by state in order to calculate this average. As we can see, there are several states that are gray, given that the dataset provided doesn't have values for them at all, which is concerning as we don't have information for 15 states. Though several of the states are varying in the green color range, something interesting to note are the states Indiana and New York, which appear to be on opposite ends of the spectrum. There are very few similarities between this map and that of PM2.5 concentrations, and it initially does not seem like the two will be related in any way.

## *Lung Cancer*

*Incidence of Cancer of the lung or bronchus, Crude Rate per 100,000*


In this visualization, we wanted to check the prevalence of lung cancer among residents of each state, across all years from 2008 to 2020. As evident from the data, the overwhelming number of states have a lung cancer prevalence rate of around ten thousand per a hundred thousand state residents. This is peculiar, as it suggests that contrary to what common sense might suggest, the average air quality of each state has little to no effect on the prevalence rate of lung cancer. The states that have deviations from the majority are mainly large population centers such as California, Texas, Florida, etc. Comparing this to the air quality map, we can also see that there are very few similarities between the prevalence of lung cancer and the trends we saw in air quality.

*COPD:*

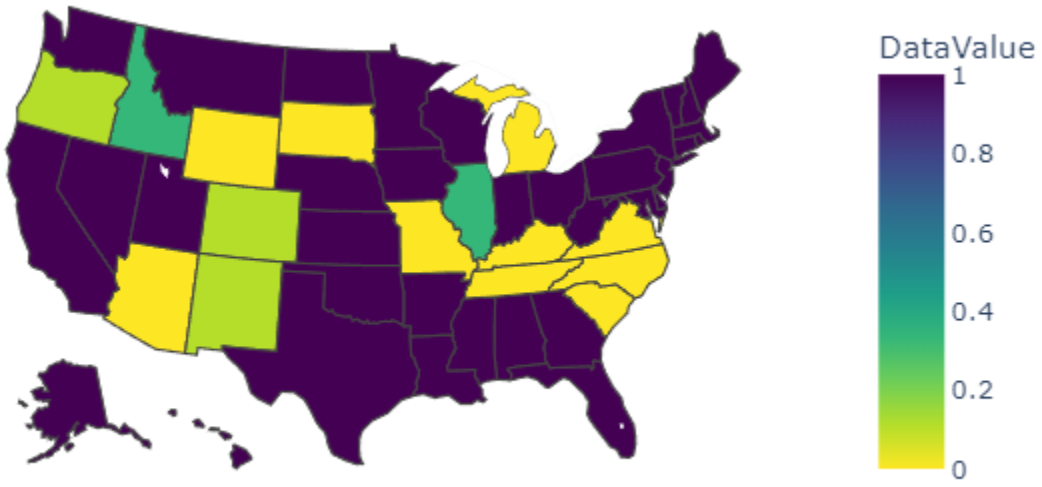



For this section, we wanted to investigate the relationship between whether a state has a higher COPD prevalence rate when compared against the national average given a specific year. For this visualization, we chose to use a Bernoulli variable, with it representing whether or not the state's prevalence rate is above or below national average.

Seeing which states lie away from the national average may in tandem with other variables such as PM 2.5 levels, ozone stratification and tobacco laws may signify possible causation, from which we would have to test individually to significantly prove such causation.

*Strong Tobacco Laws*

*Does a state have strong laws regarding tobacco sales?*




As part of our analysis, we wanted to measure the impact that tobacco usage had on the rate of lung cancer using whether or not states had strong laws regulating the sale of tobacco products as an instrument for tobacco usage. A state is determined to have strong laws if it has policy in at least one of the following categories: advertising regarding tobacco, smoke-free indoor air, and youth access to tobacco products. This was measured in the dataset across 9 different time frames, and the visualization represents in how many of those 9 time frames did the state have strong laws regarding tobacco (a score or 0.111 would mean that there were only strong laws in the most recent timeframe and no such laws existed in the first 8). For example, dark purple states had strong laws for the entire time frame, versus a state like Idaho that only

implemented strong laws after four time frames had passed and only had strong laws for the remaining five.

From the map, we can see that a majority of US States had strong laws regarding tobacco over the entire time frame. However, there are still a non insignificant number of states that have much lower scores, meaning that policies regarding tobacco were implemented only recently or not at all, which can be used to contrast with states with strong laws in our analysis.

# Multiple hypothesis testing / decision making

**Research Question 1**: *For each of the six types of diseases listed above, is there a significant association between the prevalence of the disease and the average yearly PM2.5 concentrations*

**Methods**

Our six hypothesis tests consist of whether there is a significant association between the PM2.5 concentrations and the six types of diseases listed below. Hypothesis testing was conducted on all 6 diseases/health conditions as we wanted to rule out the possibility of a disease being more prevalent due to a state's population being more susceptible to a particular disease. By testing the prevalence of all 6 conditions, we can reasonably assume that the prevalence is not caused by specific anomalies, but rather a statewide issue that affects all.

In order to accomplish this, we ran a simple linear regression for each disease and interpreted the hypothesis test on the value of the coefficient on PM2.5 concentration.

For all of the hypothesis tests,

**$H_0$:** The coefficient on PM2.5 is zero (Air quality has no effect on the likelihood of getting the disease)

**$H_A$:** The coefficient on PM2.5 is not zero (Air quality DOES have an effect on the likelihood of of getting the disease)

## *Lung Cancer*

Lung Cancer is interpreted as the average annual crude rate of the incidence of cancer of the lung and bronchus per 100,000. Below are the results of the regression.

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | y | R-squared: | 0.009 |
| Model: | OLS | Adj. R-squared: | -0.012 |
| Method: | Least Squares | F-statistic: | 0.4479 |
| Date: | Mon, 01 May 2023 | Prob (F-statistic): | 0.507 |
| Time: | 15:09:38 | Log-Likelihood: | 23.995 |
| No. Observations: | 49 | AIC: | -43.99 |
| Df Residuals: | 47 | BIC: | -40.21 |
| Df Model: | 1 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 0.4534 | 0.135 | 3.353 | 0.002 | 0.181 | 0.725 |
| x1 | 0.0101 | 0.015 | 0.669 | 0.507 | -0.020 | 0.040 |

| | | | |
|---|---|---|---|
| Omnibus: | 8.151 | Durbin-Watson: | 2.246 |
| Prob(Omnibus): | 0.017 | Jarque-Bera (JB): | 7.383 |
| Skew: | 0.764 | Prob(JB): | 0.0249 |
| Kurtosis: | 4.133 | Cond. No. | 56.7 |

The coefficient is 0.0101, meaning that with every 1μg/m3 increase in 24-hour average PM2.5 concentration, we would expect the number of people that contract lung cancer to increase by 0.0101. However, the p-value is 0.507, which means that this is not significant at the 5% significance level, and is not statistically significantly different from zero.

## *Chronic Obstructive Pulmonary Disease*

COPD is measured as the crude prevalence of chronic obstructive pulmonary disease among adults ≥ 18. Below are the results of the regression.

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | y | R-squared: | 0.043 |
| Model: | OLS | Adj. R-squared: | 0.042 |
| Method: | Least Squares | F-statistic: | 55.71 |
| Date: | Mon, 01 May 2023 | Prob (F-statistic): | 1.56e-13 |
| Time: | 15:09:38 | Log-Likelihood: | 577.92 |
| No. Observations: | 1257 | AIC: | -1152. |
| Df Residuals: | 1255 | BIC: | -1142. |
| Df Model: | 1 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 0.0474 | 0.031 | 1.545 | 0.123 | -0.013 | 0.108 |
| x1 | 0.0263 | 0.004 | 7.464 | 0.000 | 0.019 | 0.033 |

| | | | |
|---|---|---|---|
| Omnibus: | 39.602 | Durbin-Watson: | 1.643 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 47.846 |
| Skew: | 0.367 | Prob(JB): | 4.08e-11 |
| Kurtosis: | 3.612 | Cond. No. | 62.7 |

The coefficient is 0.0263, meaning that with every 1μg/m3 increase in 24-hour average PM2.5 concentration, we would expect the rate of hospitalizations for COPD per 10,000 cases to increase by 0.0263. The p-value for this test is 0, meaning this estimate is statistically significant and that there is an association between increased pm2.5 concentrations and the prevalence of COPD hospitalization cases.

## *Asthma*

Asthma is measured as the crude prevalence of asthma among adults aged ≥ 18 years. Below are the results of the regression.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.006
Model:                            OLS   Adj. R-squared:                 -0.027
Method:                 Least Squares   F-statistic:                    0.1850
Date:                Mon, 01 May 2023   Prob (F-statistic):              0.670
Time:                        15:09:39   Log-Likelihood:                 5.1070
No. Observations:                  32   AIC:                            -6.214
Df Residuals:                      30   BIC:                            -3.282
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.3830      0.265      1.447      0.158      -0.158       0.924
x1             0.0128      0.030      0.430      0.670      -0.048       0.074
==============================================================================
Omnibus:                        1.047   Durbin-Watson:                   2.035
Prob(Omnibus):                  0.593   Jarque-Bera (JB):                0.271
Skew:                           0.128   Prob(JB):                        0.873
Kurtosis:                       3.372   Cond. No.                         63.1
==============================================================================
```

The coefficient is 0.0128, meaning that with every 1μg/m3 increase in 24-hour average PM2.5 concentration, we would expect the crude rate of hospitalizations from asthma to increase by 0.0128. However, the p-value is 0.670, which means that this is not significant at the 5% significance level, and is not statistically significantly different from zero.

In the case of Asthma, the alternate hypothesis would be that there is no correlation between PM2.5 concentrations and asthma hospitalizations. In order to calculate the power of the test, we needed to find the probability of rejecting the null hypothesis given that the alternative is true. We calculated the power to be 0.05, which shows that this test is likely very inaccurate. However, we failed to reject the null hypothesis in this situation.

## *Cardiovascular Disease*

Cardiovascular disease is measured as the crude prevalence of awareness of high blood pressure among adults aged ≥ 18 years. This is likely a slight underestimation as there may be people who do not have their blood pressure checked and are unaware of their condition, but we believe that this is the best representation given the data that we are working with.

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | y | R-squared: | 0.008 |
| Model: | OLS | Adj. R-squared: | -0.013 |
| Method: | Least Squares | F-statistic: | 0.3753 |
| Date: | Mon, 01 May 2023 | Prob (F-statistic): | 0.543 |
| Time: | 15:09:39 | Log-Likelihood: | 38.644 |
| No. Observations: | 49 | AIC: | -73.29 |
| Df Residuals: | 47 | BIC: | -69.50 |
| Df Model: | 1 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 0.5829 | 0.100 | 5.814 | 0.000 | 0.381 | 0.785 |
| x1 | 0.0069 | 0.011 | 0.613 | 0.543 | -0.016 | 0.029 |

| | | | |
|---|---|---|---|
| Omnibus: | 4.202 | Durbin-Watson: | 2.080 |
| Prob(Omnibus): | 0.122 | Jarque-Bera (JB): | 3.636 |
| Skew: | 0.311 | Prob(JB): | 0.162 |
| Kurtosis: | 4.181 | Cond. No. | 56.7 |

The coefficient is 0.0069, meaning that with every 1μg/m3 increase in 24-hour average PM2.5 concentration, we would expect the crude rate of mortality from total cardiovascular diseases to increase by 0.0069. However, the p-value is 0.543, which means that this is not significant at the 5% significance level, and is not statistically significantly different from zero.

## *Mental Health*

Mental Health is measured as the prevalence of mental health issues among adults aged 18 and over. These were the results of the regression/hypothesis test.

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | y | R-squared: | 0.048 |
| Model: | OLS | Adj. R-squared: | 0.028 |
| Method: | Least Squares | F-statistic: | 2.364 |
| Date: | Mon, 01 May 2023 | Prob (F-statistic): | 0.131 |
| Time: | 15:09:40 | Log-Likelihood: | 51.483 |
| No. Observations: | 49 | AIC: | -98.97 |
| Df Residuals: | 47 | BIC: | -95.18 |
| Df Model: | 1 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 0.6800 | 0.077 | 8.813 | 0.000 | 0.525 | 0.835 |
| x1 | 0.0132 | 0.009 | 1.537 | 0.131 | -0.004 | 0.031 |

| | | | |
|---|---|---|---|
| Omnibus: | 1.874 | Durbin-Watson: | 1.982 |
| Prob(Omnibus): | 0.392 | Jarque-Bera (JB): | 1.409 |
| Skew: | 0.206 | Prob(JB): | 0.494 |
| Kurtosis: | 2.279 | Cond. No. | 56.7 |

The coefficient on PM2.5 concentrations is 0.0132, meaning that with every 1μg/m3 increase in 24-hour average PM2.5 concentration, we would expect the prevalence of mental health issues among adults aged 18 and above by 0.0132. The p-value for this test is 0.131, which when individually interpreted, means this estimate is statistically insignificant and that there is not an association between increased PM2.5 concentrations and the prevalence of mental health issues.

## *Chronic Kidney Disease*

Chronic Kidney disease is measured as a crude prevalence of chronic kidney disease among adults aged ≥ 18 years. Below are the results of the regression.

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | y | R-squared: | 0.118 |
| Model: | OLS | Adj. R-squared: | 0.099 |
| Method: | Least Squares | F-statistic: | 6.277 |
| Date: | Mon, 01 May 2023 | Prob (F-statistic): | 0.0158 |
| Time: | 15:09:40 | Log-Likelihood: | 41.992 |
| No. Observations: | 49 | AIC: | -79.98 |
| Df Residuals: | 47 | BIC: | -76.20 |
| Df Model: | 1 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 0.5187 | 0.094 | 5.539 | 0.000 | 0.330 | 0.707 |
| x1 | 0.0262 | 0.010 | 2.505 | 0.016 | 0.005 | 0.047 |

| | | | |
|---|---|---|---|
| Omnibus: | 2.247 | Durbin-Watson: | 1.853 |
| Prob(Omnibus): | 0.325 | Jarque-Bera (JB): | 2.129 |
| Skew: | 0.483 | Prob(JB): | 0.345 |
| Kurtosis: | 2.667 | Cond. No. | 56.7 |

The coefficient is 0.0262, meaning that with every 1μg/m3 increase in 24-hour average PM2.5 concentration, we would expect the prevalence of chronic kidney disease from adults to increase by 0.0262. The p-value is 0.016, which means that this is significant at the 5% significance level when interpreted individually.

## *Controlling for FWER and FDR*

In addition to these hypothesis tests, we also used the Bonferroni correction and the Benjamini Hochberg process to control for the FWER and FDR, respectively.

Our results from Bonferroni with a significance level of 0.05 and the 6 p values calculated resulted in 1 discovery: only COPD is significant. FWER measures the probability of making at least one false discovery, which in this case would be the probability of the result having a low p-value and being reported as statistically significant when its deviation was purely due to chance.

Our results from Benjamini-Hochberg with a significance level of 0.05 and the 6 p values calculated resulted in 2 discoveries: COPD and Kidney Disease are significant. FDR is the expected percentage of false positives out of all positive predictions

In our case, we believe that the results from the Benjamini-Hochberg process controlling for the false discovery rate is more relevant to our specific research question. This is because the implications behind significant test results would be to push towards better policy regulating PM2.5 concentrations and air quality to reduce likelihood of contracting diseases. Type I errors that are controlled by the Bonferroni correction aren't that important, because having a false positive result could result in positive externalities for other diseases that air quality does affect. Controlling for the FDR instead still allows for some false discoveries, but maintains them to a low enough percentage, which is more than enough for our purposes.

### *Limitations*

The most significant limitation to our analysis is the small number of p values. Since we only had 6 different hypothesis tests, making it difficult to draw conclusions. Because we have a small sample size, the power of the test is lower and there might be difficulties to detect the true effects. It is also harder to control the false positive rate.

If we had more data, we would apply hypothesis testing to the prevalence of PM2.5 to other diseases. For example, we might test whether or not PM2.5 has a link to adverse birth outcomes, such as low birth weight and preterm birth or cognitive disorders such as Alzheimer's. This would also provide a larger sample size for controlling for the FWER and FDR, which would result in more significant results.

# Causal Inference

## *Methodology*

**Treatment:** Our treatment here was the mean crude prevalence of tobacco for available years in the chronic diseases database for current smoking adults aged ≥ 18 years.

**Outcome:** Our outcome was the mean of the average annual crude rate for chronic obstructive pulmonary disease among adults aged ≥ 18 years.

There are multiple confounding variables we consider: Age, family medical history, diet and exercise are all variables that can potentially affect both treatment and outcome. This is because there are many things that can cause someone to start smoking tobacco products, and all of these confounders could potentially influence tobacco use and also risk of COPD. The unconfoundedness assumption does not hold here.

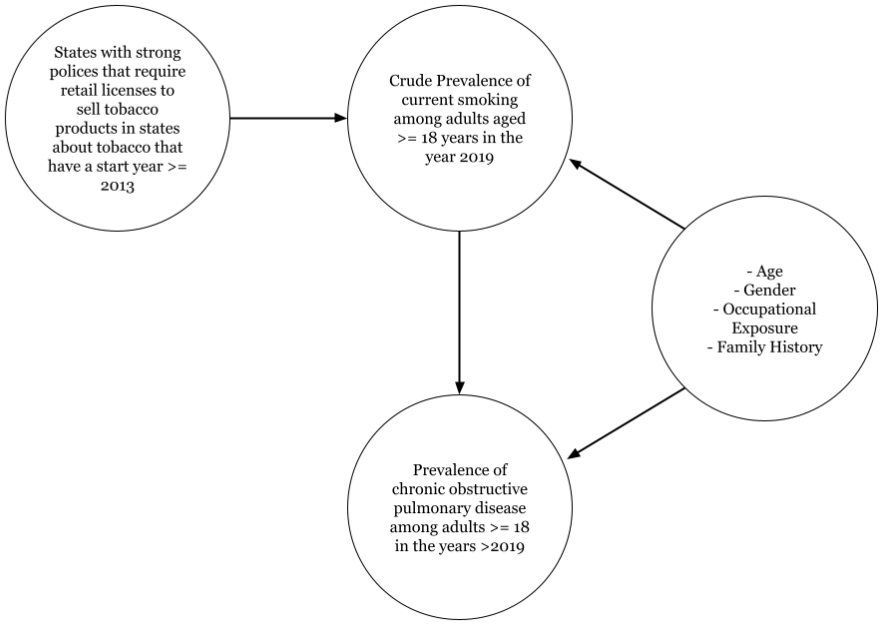
We used the two-stage least squares methods from the lab, using an instrumental variable to avoid this issue. That is, we first train our first regression model to predict our treatment using our instrument. Our instrument variable is a binary variable that determines whether or not states have strong laws regarding tobacco use.

$$PredictedTobaccoUse \sim a + \gamma_1 StrongLaws + \varepsilon \quad (Equation\ 1)$$

Then we use our predicted treatment values to train a second regression model to predict the outcome variable.

$$COPD \sim A + \beta_1 PredictedTobaccoUse + \varepsilon' \quad (Equation\ 2)$$

There are two colliders in the dataset: the treatment (tobacco consumption), and the outcome (whether or not a participant gets COPD). Below is the DAG that we used to make our model.




The data itself was processed in the following manner. For the treatment, we calculated the crude prevalence of current smoking among adults aged ≥ 18 years old in the year 2019. The instrument was whether or not the state had strong policies that require retail licenses to sell tobacco products in any year starting from and including 2013. The outcome was calculated as the prevalence of chronic obstructive pulmonary disease among adults ≥ 18 in the years after 2019. Given that we grouped by state abbreviations and aggregated by mean, the treatment values ranged from numbers between 0 to 1 given that some states had "strong laws concerning tobacco" in some years and not in others. We changed the values for the "Strong Laws" column

to 0 if the value was 0 since this means that the state never had such laws starting from 2013 and 1 if it was > 0, as the state itself had the law at some point after and including 2013.

### *Assumptions*

There are two major assumptions that need to be made for our instrumental variable to be legitimate: exogeneity and the exclusion restriction. Exogeneity means that the instrument must be uncorrelated with the error term in the regression. We are assuming this to be true because it is highly unlikely that state laws are related to any of the confounding variables omitted from the model that we called out, and thus it is exogenous from the error term.

Second, state laws need to only affect the outcome variable (COPD prevalence) through its impact on the explanatory variable (Tobacco use). This we also assume to be true, as it is highly unlikely that laws regarding tobacco sales would affect anything else besides tobacco use, as the laws are extremely specific and niche.

### *Results*

The results of the first stage regression (Equation 1), is the following.

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | Tobacco Crude Prevalence | R-squared: | 0.115 |
| Model: | OLS | Adj. R-squared: | 0.097 |
| Method: | Least Squares | F-statistic: | 6.466 |
| Date: | Mon, 08 May 2023 | Prob (F-statistic): | 0.0141 |
| Time: | 21:36:04 | Log-Likelihood: | -147.22 |
| No. Observations: | 52 | AIC: | 298.4 |
| Df Residuals: | 50 | BIC: | 302.3 |
| Df Model: | 1 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 20.9239 | 1.324 | 15.806 | 0.000 | 18.265 | 23.583 |
| Strong Laws | -3.7454 | 1.473 | -2.543 | 0.014 | -6.704 | -0.787 |

| | | | |
|---|---|---|---|
| Omnibus: | 1.261 | Durbin-Watson: | 1.441 |
| Prob(Omnibus): | 0.532 | Jarque-Bera (JB): | 1.127 |
| Skew: | 0.183 | Prob(JB): | 0.569 |
| Kurtosis: | 2.379 | Cond. No. | 4.36 |

The coefficient on strong laws is -3.7454. This aligns with expectations, as we would expect states with stronger laws to also have lower rates of tobacco usage. Furthemore, the p-value is 0.014, which is statistically significant at the 5% significance level. This satisfies the "relevance" condition of an instrumental variable — strong laws are strongly correlated with tobacco use. The instrumental variable also only affects the outcome through its effect on the treatment, meaning it is a valid instrumental variable.

The results of the second stage regression (Equation 2) is below:

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | COPD Crude Prevalence | R-squared: | 0.064 |
| Model: | OLS | Adj. R-squared: | 0.045 |
| Method: | Least Squares | F-statistic: | 3.417 |
| Date: | Mon, 08 May 2023 | Prob (F-statistic): | 0.0705 |
| Time: | 22:02:23 | Log-Likelihood: | -108.62 |
| No. Observations: | 52 | AIC: | 221.2 |
| Df Residuals: | 50 | BIC: | 225.1 |
| Df Model: | 1 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 0.4387 | 3.362 | 0.130 | 0.897 | -6.314 | 7.192 |
| Predicted Tobacco Crude Prevalence | 0.3460 | 0.187 | 1.848 | 0.070 | -0.030 | 0.722 |

| | | | |
|---|---|---|---|
| Omnibus: | 14.695 | Durbin-Watson: | 1.985 |
| Prob(Omnibus): | 0.001 | Jarque-Bera (JB): | 16.610 |
| Skew: | 1.138 | Prob(JB): | 0.000247 |
| Kurtosis: | 4.576 | Cond. No. | 219. |

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

The coefficient for the predicted tobacco crude prevalence from our stage 1 regression is 0.3460, which means that if the prevalence of tobacco use were to increase by 1%, we would expect to see an increase of 0.0034% prevalence of COPD in the same state. The p-value for this estimate is 0.07, which is significant at the 10% significance level but not at the 5% significance level that is standard and the threshold for all other tests conducted in this paper.

Although we expected that there would be detrimental effects of consuming tobacco and COPD, according to our conclusion, that is not the case. There have been many studies that have shown the link between tobacco use and COPD, such as the study done by the American Lung Foundation which found that 3 out of 4 COPD patients have a history of smoking.

After conducting our analysis, we found little evidence to support a correlation between COPD prevalence and tobacco use prevalence. However, it is important to acknowledge that our conclusion is subject to certain limitations stemming from the dataset at hand. If we had access to more relevant data, we may have arrived at different conclusions, and perhaps enough of a difference would have been made for the result to be significant at the 5% significance level as well.

To ensure the accuracy of our causal inference, we followed established best practices, such as controlling for confounding variables and accounting for temporality. Specifically, we used a two-stage least squares method to control for potential confounding variables and establish an unbiased connection between the treatment (prevalence of tobacco users) and outcome (COPD prevalence). We also accounted for temporality by ensuring that the outcome was not calculated before the treatment, thereby avoiding any false connections between non-smokers and COPD. (what does this last sentence mean -alvin)

## *Discussion and Limitations*

Below are some limitations on our methods:

- ☐ *Selection bias*: our tobacco users are already within the chronic disease dataset.
- ☐ *Measurement bias*: Smokers lied about being non-smokers.
- ☐ *Generalizability*: Our conclusions may not be generalizable to other populations or settings.
- ☐ *Incorrect instrument variable* : There is potential association between our instrument variable and our response.
- ☐ *Not being able to perform randomization:* Due to our causal inference working with data points that are each tied to a specific location, we're not able to effectively perform randomization before the inference process

Our dataset for participants with lung cancer only included annual crude prevalence rate, which is helpful to an extent, but very barebones to transform the association between tobacco consumption and COPD prevalence into causation. Some helpful data that would be useful in answering this causal question would have been the age and gender of the participant, their occupation and what they are exposed to, and their family medical history. Age and gender would be useful due to COPD being prevalent amongst elderly, and females are more at risk of developing COPD. Their occupation might give more insight on what kind of chemicals they are exposed to that might lead to COPD Additionally, numerous studies have demonstrated that genetic factors also contribute to the development of COPD. By knowing all the additional factors, we can more conclusively determine the effects of tobacco consumption.

# Conclusions

## *Hypothesis testing*

For the hypothesis testing section, we made multiple discoveries while correcting for both FPR and FWER. When correcting using Bonferroni, we found one discovery( COPD), while when correcting using Benjamini-Hochberg, we found that COPD and Kidney Disease are significant.

That is, PM.2.5 has a fairly high chance of having detrimental effects to these diseases. However, we also realized we had a limited number of p-values, which could potentially increase the likelihood of type 1 errors. Future work could include adding more tests to our multiple hypothesis testing, or diving deeper into different types of pollutants. It is highly likely that PM2.5 could have associations with many other diseases that weren't in our tests. With this in mind, it is imperative that we try our best to reduce PM2.5 pollution. For example, encouraging individuals to make lifestyle changes such as using public transportation or raising awareness about the potential risks of PM2.5 pollution.

## *Causal Inference*

For the causal inference section, we wanted to measure the causal relationship between tobacco consumption and COPD This was done by using an instrument variable (Strong laws of a state) and two stage least squares regression. We discovered a negative causal effect between the two variables as expected, as we would expect stronger laws to reduce tobacco consumption. However, we must also consider the limitations of our methods. There could have been potential errors and biased samples that could have steered our results. For future work, we could redo this experiment with other causal inference methods such as matching or inverse propensity

weighting. If we were to make a call to action based on similar studies that were conducted (association between tobacco consumption and COPD), it would simply be to discourage tobacco use.