Modeling U.S. Asthma Prevalence and Particulate Matter 2.5 Concentrations

Data 102 Final Report: Razi Mahmood, Carlos Ortiz, Aileen Peralta, Isabel Zavian

Data Overview

The project datasets come from the Centers for Disease Control and Prevention (CDC). We used two subsets of the "U.S. Chronic Disease Indicators" dataset, which contains state-specific data from 2010 to 2019 for asthma and tobacco use prevalence, and a sampled subset of the "Daily Census Tract-Level PM2.5 Concentrations" dataset, which records 24-hr Avg. PM2.5 concentrations at a census tract-level for the U.S. from 2011 to 2014. The datasets come from a credible source, so we assume their data to be accurate, reliable, and up-to-date.

The "U.S. Chronic Disease Indicators" subsets have a structure where each row represents asthma/tobacco use prevalence for adults 18 and older with respect to state and year. The data comes from nine sources, of which the CDC cites as having "complex sample designs and weights," so it's uncertain as to whether or not the individuals are aware of this data collection. There are concerns of measurement error that relates to the spatial contexts of communities; some groups are systematically excluded, namely BIPOC, immigrant, undocumented, and more. This will impact the interpretation of our findings as it will skew towards making sense of individuals who have access to healthcare and are able to participate in studies. For example, the subset of asthma prevalence has null values for 26 states in 2014 at a race/ethnicity stratification of Hispanic. We discuss this limitation later in the report.

The "Daily Census Tract-Level PM2.5 Concentrations" dataset has a structure where each row represents a 24-hr Avg. PM2.5 concentration with respect to day, county, and state. There are concerns of selection bias in our data since it comes from sensor readings that vary in location and quantity. PM2.5 varies within counties, influenced by land cover, so a sensor's placement in a park can produce different readings compared to a sensor's placement on a busy street next to it. We aggregate the data to produce averages, but the underlying limitations will still impact the interpretation of our findings, which we discuss later in the report.

Research Questions

Our first research question, "Does the data for asthma prevalence and 24-hr Avg. PM2.5 concentrations differ at different stratification and scales?" is examined using multiple hypothesis testing. It allows us to understand the underlying distributions and relationships of the data. We explore this with these five questions:

1. Is there a difference in U.S. states' asthma rates from 2011 to 2014?
2. Is there a difference in U.S. states' asthma rates across males and females?
3. Is there a difference in U.S. states' asthma rates across White, Non-Hispanic and Hispanic?
4. Is there a difference in county 24-hr average PM2.5 concentrations across LA and Alameda County?
5. Is there a difference in state 24-hr average PM2.5 concentrations across California and Texas?

For each hypothesis test, the null hypothesis is that there is no significant difference between the two distributions and any difference is due to chance; the alternative hypothesis is that there is a significant difference. These hypothesis tests are more resource allocation questions, as in, if there are differences at different stratification levels for asthma prevalence, then it should be accounted for in policy; likewise for PM2.5, if there are differences, some of it could be attributed to human/non-human causes. Multiple hypothesis testing works for the principal question as it aims to account for the type I error in just one hypothesis test.

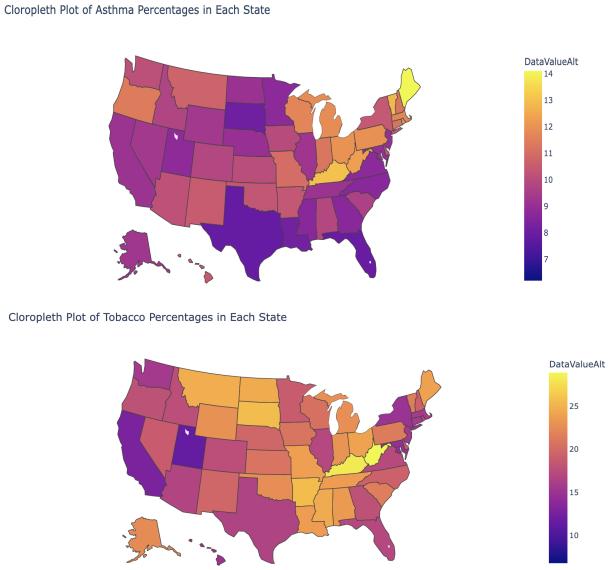
Our second research question, "Do higher levels of PM2.5 cause higher levels of asthma?" is examined using causal inference. This could motivate policy, especially as the PM2.5 data is at a census tract-level and could motivate directed policy that supports communities from the effects of PM2.5 among other air pollutants. Causal inference is a good fit for the question as it aims to quantify the effect of PM2.5 on asthma.

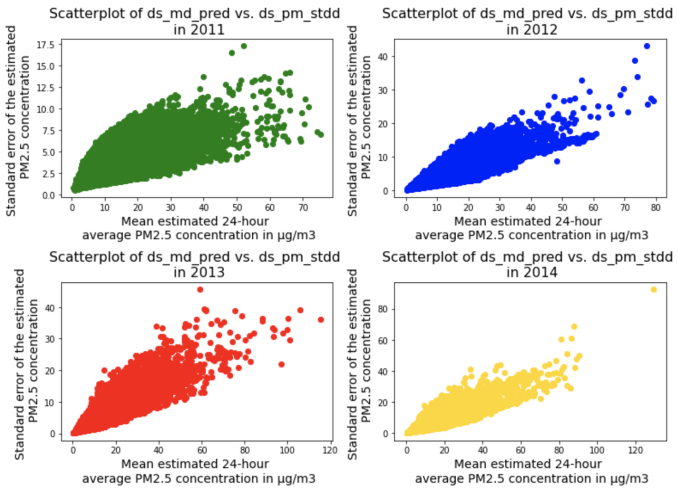
Exploratory Data Analysis

The focus for EDA is to look into the relationship between asthma/tobacco use prevalence and 24-hr Avg. PM2.5 concentrations at different stratifications and aggregations.

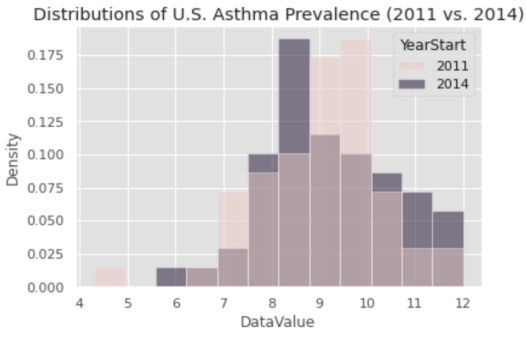
In the initial step of data cleaning, we first examined what the indicator dataset captures. We had to consider how the data comes from different sources, is aggregated by year and state, and has different units the values are measured in. We filtered the data to prevalence among adults age 18 and older, and averaged across sources for each year and state. Then, we split it into different stratifications to capture gender, race/ethnicity, and time. For the PM2.5 dataset, we aggregated to two different scopes: county-level and state-level. For example, if we needed data for California, then we only selected the rows that contained a statefip of '6.' To narrow it down even further and look at the PM2.5 for LA County, we filtered the data to only select rows that contained a countyfip of '6037.' To merge the indicator data to PM2.5 data, we grouped them at a state-level, aggregated by average.

For the causal inference question, we are looking into whether higher concentrations of PM2.5 cause higher levels of asthma. There are possible confounding variables, so our EDA for this consists of two quantitative choropleth maps—one for the asthma indicator and another for the tobacco use indicator by the granularity of state. It is possible that tobacco use could have an impact on asthma levels for our research question, so we examine the relationship between both. As shown in Figure 1, there isn't much of a visual correction between the two, so to quantify this, we calculated the correlation. We got a correlation of -0.17 which is fairly weak, so we found that possible confounding factors, such as tobacco use, won't have a significant influence in our research study.

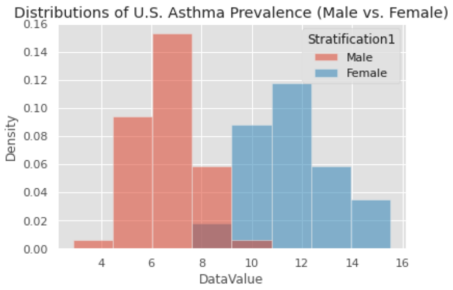


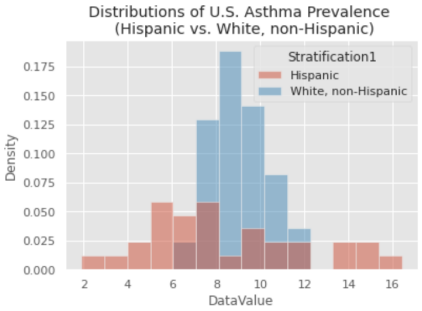


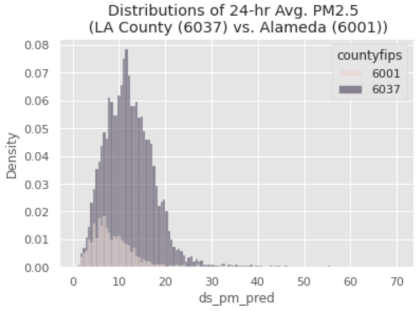
We also plotted multiple scatterplots to identify the relationship between the mean estimated 24 hour average of the PM 2.5 concentration and the standard deviation of the PM 2.5 concentration for each year between 2011-2014. We noticed that some of the plots had slight differences in their distributions, which could've been due to outliers in the randomly sampled data, but all four plots indicated a positive correlation between the two variables. Because of this, we included the standard deviation of the PM 2.5 levels as one of our treatment variables for causal inference.

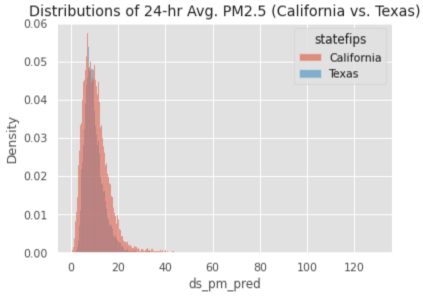


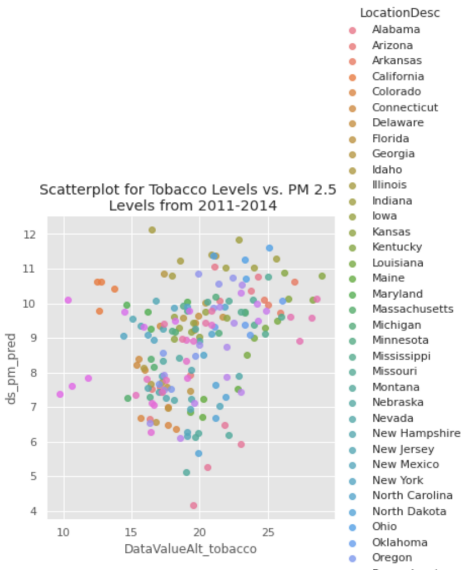
Since we came up with five different hypothesis questions, we plotted a variety of histograms pertaining to each one to observe the distributions of specific components of the datasets. The histogram on the right shows the distributions of asthma prevalence in adults 18 years and older in 2011 and 2014. The underlying distributions are the same, but there is a slight shift down in the average asthma prevalence values from 2011 to 2014.

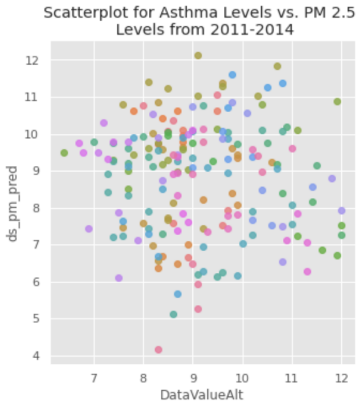


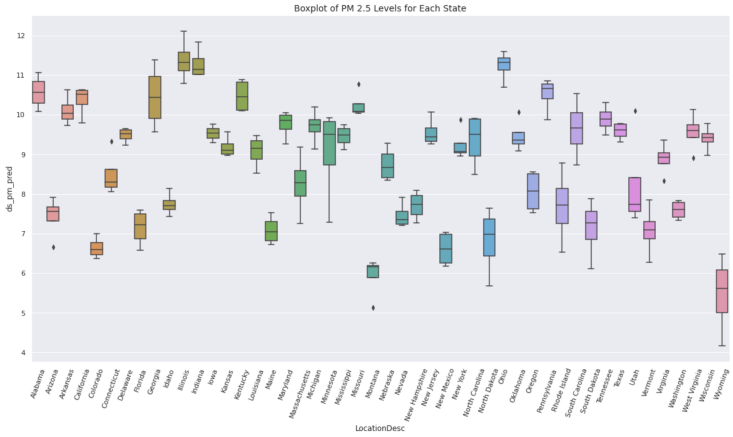








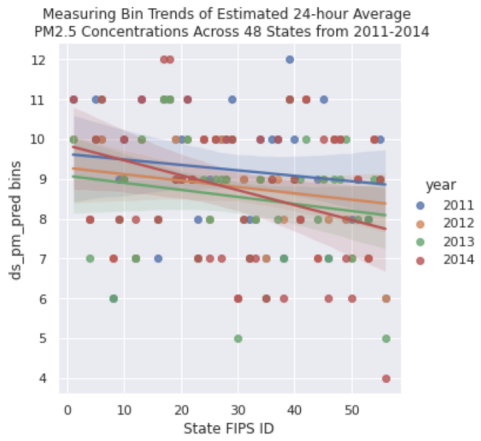


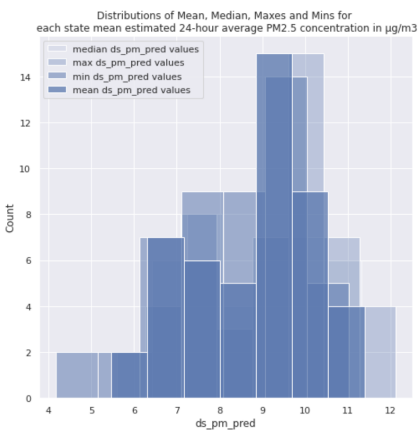


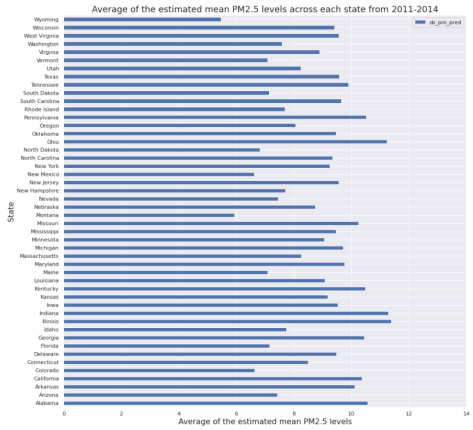
To the left are a couple more of the histograms we generated to observe the distributions of U.S. asthma prevalence between males and females, asthma prevalence between Hispanics and non-Hispanics, and the 24 hour average PM2.5 levels between LA County vs. Alameda and California vs. Texas in 2014. These histograms suggest that there are differences in the distributions for gender in asthma, and county for PM2.5. This motivates our multiple hypothesis testing.

The two scatterplots attached below show the relationship between the asthma levels and PM2.5 levels between 2011-2014 and the tobacco levels and PM2.5 levels between 2011-2014. Each of the 48 states have 4 points on both plots because every point represents the average asthma or tobacco levels for that given year. Although it's difficult to distinguish between some of the colors and find a relationship between the features, there isn't much of a correlation for all the data in general. We took a deeper look at this by creating a separate plot for each state with all 4 points.

We also created a boxplot to observe the range of PM2.5 levels for every state across each of those years. Moreover, we set bins ranging from 4-12 for the estimated 24-hour average PM2.5 concentrations across all 48 of the states between 2011-2014. This was done to identify the overall distribution of the concentrations for not only each state but each year overall (as shown in the slopes in the right diagram). Another useful visual was a histogram representing the mean, median, max, and min for the asthma prevalence percentages in the entire dataset without focusing on specific states or years. Lastly, a barplot displaying the average estimated mean of PM2.5 levels for each state from all four years combined helped identify overall patterns in which states had higher concentrations in general.







Methods, Inference and Decisions

For multiple hypothesis testing, we ask five hypotheses; three relate to different stratifications of U.S. states' asthma prevalence for adults age 18 and older (i.e., time, gender, race/ethnicity), and two relate to different scales of 24-hr Avg. PM2.5 rates (i.e., county-level and state-level). It makes sense to test many hypotheses compared to just one because different stratifications and/or scales will impact our decisions, especially as our threshold impacts our type I error. Each of the five hypotheses are tested using A/B testing to break down our distributions. The first three tests examine how each stratification contributes to the overall asthma prevalence and the last two examine how 24-hr Avg. PM2.5 concentrations vary by location. These questions are important as they further contextualize our understanding of the relationship between PM2.5 rates and asthma prevalence. We make the following decisions:

1. There is no significant difference in U.S. states' asthma prevalence between 2011 and 2014. Any difference is due to chance. We fail to reject the null.
2. There is a significant difference in U.S. states' asthma prevalence between males and females in 2014. We reject the null.
3. There is no significant difference in U.S. states' asthma prevalence between White, Non-Hispanic and Hispanic. Any difference is due to chance. We fail to reject the null.
4. There is a significant difference in 24-hr Avg. PM2.5 concentrations across LA County and Alameda County in 2014. We reject the null.
5. There is a significant difference in 24-hr Avg. PM2.5 concentrations across California and Texas in 2014. We reject the null.

We applied the correction procedures of Bonferroni and Benjamini-Hochberg, controlling for FWER and FDR respectively. The same discoveries remain significant in both procedures. This is likely due to the fact that their p-values were 0, which strongly suggests that the original samples do not have the same underlying distribution.

There are some limitations to our analysis. For the U.S. states' asthma prevalence across gender and race/ethnicity hypothesis tests, there is missing data for some states in 2014. This is most significant in the race/ethnicity test, where 3 states did not have data on White, Non-Hispanic asthma prevalence compared to 26 states which did not have data on Hispanic prevalence. This is especially important to consider how this factors into the other asthma prevalence hypothesis tests. For the 24-hr Avg. PM2.5 concentration hypothesis tests, we must consider how the data is significantly aggregated. Air pollution varies from block-to-block, impacted by location, land cover, and more. Although it makes sense for a city or county to have an averaged PM2.5 that communities are exposed to, this just isn't the case. This understanding is most important for the hypothesis test that compares averaged PM2.5 by state; at this point, it is too aggregated.

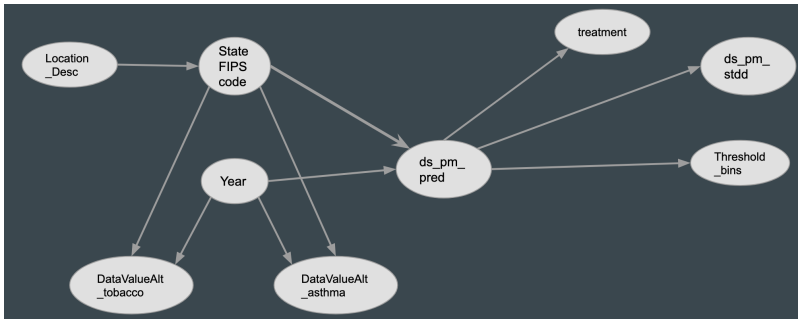
We avoided p-hacking by writing our questions first and sticking to the same sampling and statistical procedures. For example, for the asthma prevalence, we take the absolute difference between the averaged asthma prevalence from both distributions; for the 24-hr Avg. PM2.5 concentrations, we take the absolute difference between the 50th percentile from both distributions. We chose average for asthma prevalence as the distributions appear approximately normal, but chose 50th percentile for 24-hr Avg. PM2.5 concentration due to the distributions having a significant right-skew caused by outliers—likely due to weather and environmental conditions such as wildfire season. Lastly, if we had additional data, we would conduct additional hypothesis tests across different stratifications for asthma prevalence in adults 18 and older, as well as additional hypothesis tests across more years beyond 2011-2014 for PM2.5.

For our causal inference question, we ask: Do higher levels of PM2.5 cause higher levels of asthma? We attempted this question by examining possible correlations between the reported average 24-hour PM2.5 concentrations per year between 2011 and 2014 and the data values representing the asthma prevalence levels in adults aged at least 18 years from 2011-2014.

Our experiments focused on cleaning the dataset with the states and years that have overlap between the asthma and PM2.5 datasets, and applying the same transformations onto the tobacco dataset to add more covariates. Our intended treatment variable is ‘treatment’ which comes from using the ‘ds_pm_pred’ values (estimated 24-hr Avg. PM2.5 concentration levels) to distinguish between higher levels of PM 2.5 by using a threshold from our observed trends and the CDC average of PM2.5 levels from 2011-2014. The intended outcome variable is ‘DataValueAlt_asthma’ representing the asthma prevalence levels for adults 18 and older. We included covariates that are relevant to the causal inference question, such as: ‘DataValueAlt_tobacco’ representing the percentage of tobacco use prevalence, ‘statefips’ and ‘LocationDesc’ to represent the code for a specified US state, ‘ds_pm_stdd’ representing the standard error of the estimated PM2.5 concentration, and ‘year’ ranging from 2011 to 2014.

While merging the dataset, some confounders we considered including in our models were county fips, latitude and longitude. However, we could not include them because we aggregated the data by state level due to asthma and tobacco dataset only having data at state level. By not including countyfips, we are unable to analyze the effects of PM 2.5 by county. We decided to focus on getting the average of the PM2.5 estimates for a state in a given year from 2011-2014 to get past this issue. Thus, the countyfips and latitude/longitude variables are confounding variables that we excluded that affect the aggregation of the PM2.5 values and variability of the asthma prevalence values. This also relates to the Simpson’s Paradox of the confounder affecting our ‘treatment decision’ (which we used our ds_pm_pred values to compare to a threshold of 9.5) and our outcome variable of asthma prevalence.

However, confounders that we had included based on the Causal DAG are the ‘statefips’ and ‘year’ variables because the state location in both the PM2.5 and indicators datasets play a factor in how the asthma prevalence levels and estimated PM2.5 averages were distributed across each year per state. When merging the data, we had to exclude the years that were not 2011-2014 and regions, such as Hawaii or Alaska, that were present in one dataset but not in the other. These selections limited the number of data points we would receive per each state, along with the number of states we were measuring the PM2.5 levels that we may not be accounting for due to limitations in the datasets are weather forecasts and climate waves affecting the spread of the PM2.5 pollution rates. Different states may also have a varying amount of resources to counteract potential hazards from PM2.5 pollution and treat patients with asthma, or different counties for a state may have better tools than others for assistance. The limitations of



relevant covariates being filtered and merged between the 2 datasets have narrowed our choices for relevant confounders across time and location. Thus, we believe we've accounted for as many confounding factors we could find to work with an unconfoundedness assumption.

For colliders, we were having trouble finding variables within the datasets that were being influenced by the treatment/ds_pm_pred values and DataValuealt_asthma variables. Since DataValueAlt_asthma is the target variable measured in the asthma indicators set, there doesn't appear to be any specific covariates that are affected by it. The latitude-longitude variables in the PM2.5 set wouldn't work for a similar reason as the countyfips since they are both variables simply measuring a location and wouldn't be influenced by a difference in asthma or PM2.5 values. No colliders have been discovered with our merged data.

To adjust for the confounding variables, we considered the outcome regression technique since we were working with an unconfoundedness assumption and attempted to fit an OLS model to account for relevant covariates across our merged dataset. This was to help us check if the mean squared error and treatment coefficient would indicate a possible linear interaction between our treatment and outcome variables on a linear model, given the selected covariates. We also implemented the inverse propensity weighting and propensity scores from the treatment variable made based on the ds_pm_pred values to see if there is a significant treatment effect between high and low PM2.5 values conditioning on the covariates. We couldn't use exact matching because the data values for the PM2.5 dataset were aggregated across each state per year in 2011-2014, which would impact the exact conditions of the relevant confounding factors we're looking for in the asthma set.

In order to estimate the causal effect of year, statefips, ds_pm_stdd, and DataValueAlt_tobacco on DataValueAlt_asthma, we ran an OLS model. From our model's summary, we got an R2 of 0.036. The estimated causal effect on DataValueAlt_asthma of an additional PM 2.5 unit is 0.089. The next step was to estimate causal effects using the unconfoundedness assumption by converting ds_pm_pred to a binary treatment. In order to convert ds_pm_pred into a binary treatment, we needed to decide on a threshold on who received the treatment and who didn't. Based on visuals of the mean, median, max and min for each state's ds_pm_pred average, along with the reported CDC average from 2011-2014, we chose 9.5 as the best threshold to use since the distributions all had a peak close to 9-10. We calculated the simple difference in the observed group means and got -0.16, which is negative because of confoundedness. The two groups are different as individuals in treatment (states with PM2.5 >= 9.5) might face more disadvantages than the non-treatment. Because we are looking at an observational study, we have to deal with confounders and need to make the assumption of unconfoundedness. We perform an OLS model including the treatment variable and not including ds_pm_pred. We assume that the explanatory variables represent all confounding variables used in our OLS model. For our OLS model, we also assume this linear model correctly describes the interaction between the variables. We ran the OLS model with no intercept and calculated the mean squared error and got a value of 1.49. The next step was inverse propensity weighting, and we computed the propensity scores using a logistic regression model. We calculated the IPW estimate to be 1.07 and interpreted that the treatment did have a beneficial causal effect on the outcome because the estimated treatment effect is positive. One possible uncertainty in our estimates that affect our results is the threshold we chose on who gets the treatment. Although we chose 9.5 based on CDC and our visuals, results would vary with other values, which would also affect our OLS model. Another uncertainty is since we took the average PM2.5 for each state, we didn't account for county PM2.5 which could have led to different treatment groups depending on what counties had a PM2.5 greater than the threshold.

As mentioned earlier, we had to filter the years only across 2011-2014 since that was the only overlapping time period between the indicator and PM2.5 sets, giving us only 4 data points for the filtered states present in both the datasets. Along with the smaller subset of the PM2.5 dataset we had to work with since the original size was very large to upload, there were a large number of variables in the indicators set that were either blank or redundant, which made it difficult to understand the relevance. Some of the variables that did seem relevant were organized differently for each dataset since the PM2.5 set was arranged by county per state

on specific days from 2011-2014 accounting for specific latitude-longitude values, while the asthma was only arranged by state from 2010-2019. The only major variable that was common between the 2 datasets was 'LocationDesc', which could be mapped to the unique statefips code. Each row for the asthma set was answering a specific 'stratification' question, so our focus on adults at least 18 years old limited the number of prevalence values when merging. The biggest limitation came from the PM2.5 aggregation since we only had 4 data points per state representing the average PM2.5 levels per year in 2011-2014, which isn't sufficient data to measure a clear trend or properly estimate a clear causal relationship between PM2.5 and asthma.

Some additional data would be to have asthma data at the county level to do more analyses with the original non-aggregated PM2.5 dataset. Another useful approach is to include more years for the PM2.5 dataset since we only have data for 4 years, while the asthma dataset has 9 years. This would allow us to have more data per state and be more helpful in seeing if there's a linear relationship. Additionally, keeping track of resource allocation across different counties in medical & healthcare departments would be helpful to understand the frequency of patients receiving proper care for asthma and how much planning is required to prevent a hazardous spread of PM2.5 pollution. This can be done through measurements of sensors to detect accurate levels of concentration and the average spread of PM2.5 across different climate/weather conditions. Resource allocation can be a confounder since the percentage of adults diagnosed with asthma due to high pollution levels or air quality can be affected by this covariate.

From our results, we can't say there is a perfect causal relationship between treatment and outcome. However, we saw that the treatment did have an effect on the outcome because of the positive IPW estimate. We got an IPW estimate of 1.07, but we don't know what the true estimate is to make a comparison. There could be other impactful confounding variables that we didn't account for in the model to adjust the IPW estimate. We also don't have enough information about the significance of the confounding variables for the estimates or how much impact they may have in a real-world setting. For the OLS model including the treatment variable, we got a high mean squared error, so our confounders might not have a linear relationship to DataValueAlt_asthma, but there is a possibility for some type of nonlinear relationship to exist.

Conclusion:

In this report we examine the U.S. asthma prevalence and 24-hr Avg. PM2.5 concentrations through multiple hypothesis testing and causal inference. Through multiple hypothesis testing, we decide that stratification on gender results in two different underlying distributions for U.S. asthma prevalence in 2014, as well as decide that county-level and state-level distributions of 24-hr Avg. PM2.5 concentrations are different. Through causal inference, we find that there is no clear significant causal effect of PM2.5 concentrations on U.S. asthma prevalence; there isn't enough information for our analysis. In light of these findings, we call for directed policies that support underserved populations most affected by air pollution at a census tract-level. This works to address specific community needs as we were able to determine through our findings in multiple hypothesis testing.

Our results are rather generalizable and answer broad questions since we merge CDC data for U.S. Disease Indicators and 24-hr Avg. PM2.5 concentrations at very different granularities. One benefit of merging is that we can perform analyses comparing PM 2.5 and asthma. A consequence of merging the dataset is that we can only look at state level due to granularity of asthma. Limitations that we did not account for in the data are the faithfulness of it. Although our data comes from the Centers for Disease and Control (CDC), it is actually collected from 9 primary public and private sources for the disease indicators. This could influence our decisions and findings strongly, yet it isn't something we could account for in our analysis.

Future studies could build on this work by examining, and quantifying, the causal effect of 24-hr Avg. PM2.5 concentrations on different respiratory diseases; this includes asthma, acute respiratory distress, chronic obstructive pulmonary disease, lung cancer, and others in different contexts. For example, this report presents a framework in a U.S. context, which could vary greatly from other countries or territories.

References

- “Particulate Matter (PM2.5) Trends.” EPA, Environmental Protection Agency, 26 May 2021, www.epa.gov/air-trends/particulate-matter-pm25-trends.