Data C102 Final Report: Relationship Between Chronic Illnesses and Air Pollution

Emily Lopez, Maile Mayer, Noor-Ul-Ain Ali

Introduction

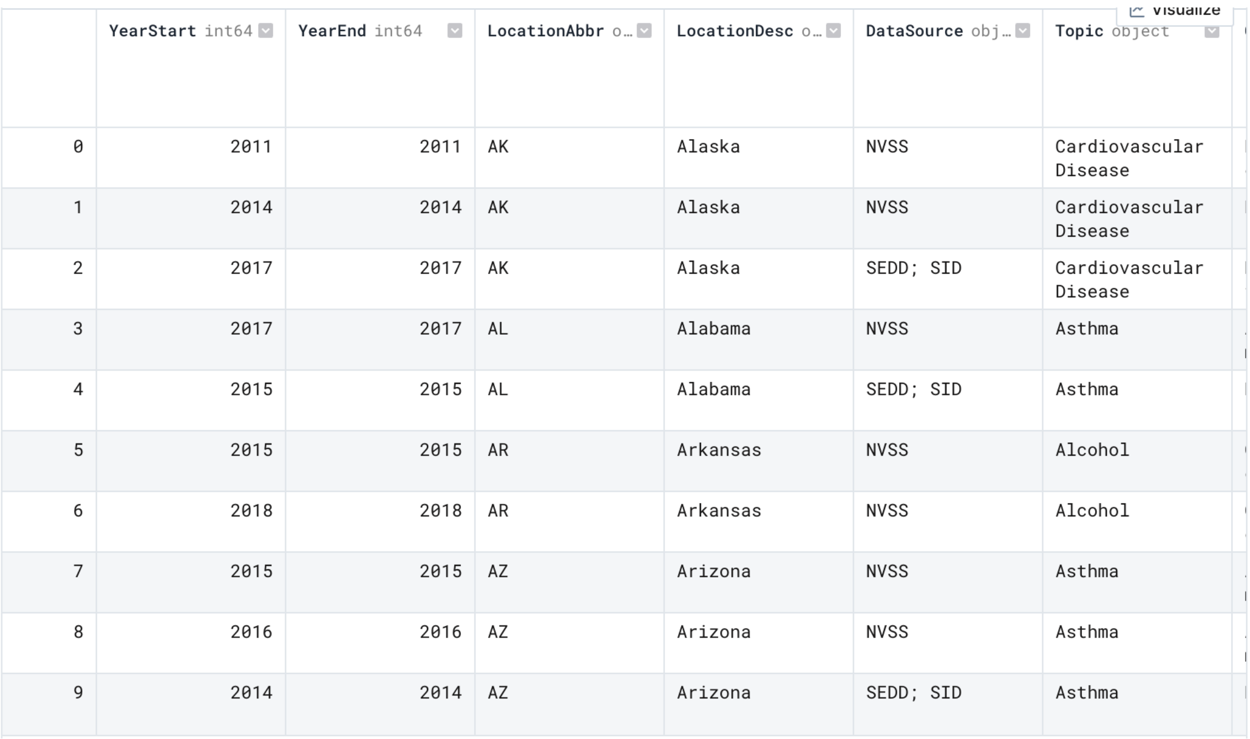
With the rise of the COVID-19 pandemic, we have seen how individuals who suffer from chronic illnesses are more susceptible to serious implications from the infection than those without. Moreso, chronic illnesses account for seven of the ten leading causes of death in the United States, three of which are considered preventable deaths, including death from chronic tobacco use and alcohol consumption.[1]

Current events have shed light on the severity of chronic illnesses, and we believe it is important to look for strategies to prevent these diseases. This may be achieved by analyzing how air pollution and chronic illnesses are related. In our research, we specifically aim to answer the questions "do increased levels of air pollution correspond with increased levels of Chronic Obstructive Pulmonary Disease (COPD)?" and "do increased regulations correspond with decreased chronic illnesses?"

1. Data Overview

Our research required the use of three datasets, two of which were combined.

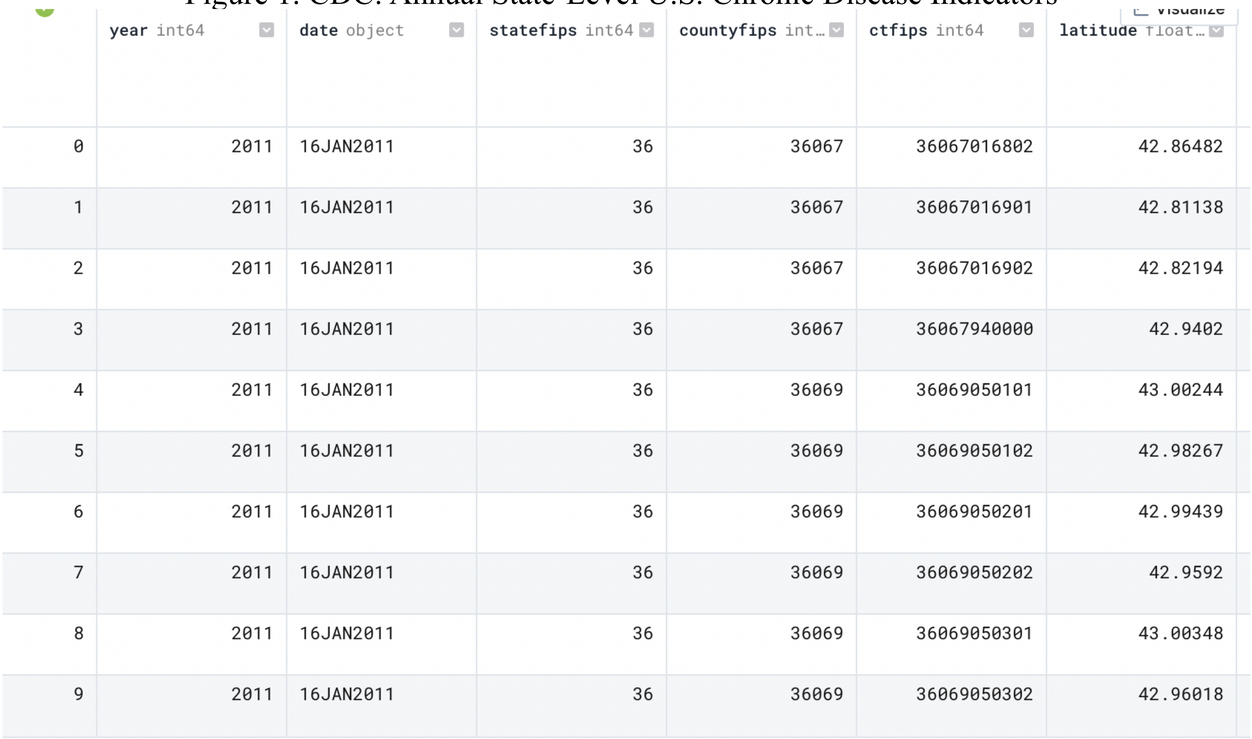
Dataset 1, Chronic Disease and Air Quality, contains data that was generated in multiple CSV datasets: (1) CDC: Annual State-Level U.S. Chronic Disease Indicators (2) CDC: Daily Census-Tract PM2.5 Concentrations. The PM2.5 concentration data was broken down by year. However, to create one dataset for our analysis, we used the yearly average for PM2.5 concentrations and combined it with the CDC dataset on chronic illnesses.



	YearStart int64	YearEnd int64	LocationAbbr o…	LocationDesc o…	DataSource obj…	Topic object
0	2011	2011	AK	Alaska	NVSS	Cardiovascular Disease
1	2014	2014	AK	Alaska	NVSS	Cardiovascular Disease
2	2017	2017	AK	Alaska	SEDD; SID	Cardiovascular Disease
3	2017	2017	AL	Alabama	NVSS	Asthma
4	2015	2015	AL	Alabama	SEDD; SID	Asthma
5	2015	2015	AR	Arkansas	NVSS	Alcohol
6	2018	2018	AR	Arkansas	NVSS	Alcohol
7	2015	2015	AZ	Arizona	NVSS	Asthma
8	2016	2016	AZ	Arizona	NVSS	Asthma
9	2014	2014	AZ	Arizona	SEDD; SID	Asthma

[1] https://www.cdc.gov/mmwr/preview/mmwrhtml/rr5311a1.htm

Figure 1. CDC: Annual State-Level U.S. Chronic Disease Indicators

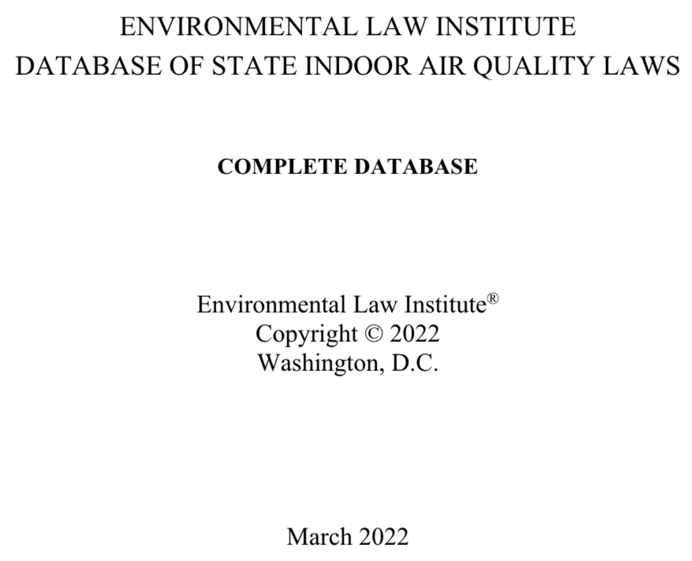


	year int64	date object	statefips int64	countyfips int…	ctfips int64	latitude float…
0	2011	16JAN2011	36	36067	36067016802	42.86482
1	2011	16JAN2011	36	36067	36067016901	42.81138
2	2011	16JAN2011	36	36067	36067016902	42.82194
3	2011	16JAN2011	36	36067	36067940000	42.9402
4	2011	16JAN2011	36	36069	36069050101	43.00244
5	2011	16JAN2011	36	36069	36069050102	42.98267
6	2011	16JAN2011	36	36069	36069050201	42.99439
7	2011	16JAN2011	36	36069	36069050202	42.9592
8	2011	16JAN2011	36	36069	36069050301	43.00348
9	2011	16JAN2011	36	36069	36069050302	42.96018

Figure 2. CDC: Daily Census-Tract PM2.5 Concentrations 2011

We also referred to ELI Database of State IAQ Laws[2] which gave us access to the various state policies that have been implemented to improve indoor air quality. Our research question "Do increased regulations correspond with decreased chronic illnesses?" involved looking at regulations at the state level, therefore, this dataset provided the necessary information. Using a dictionary of state abbreviations, we were able to scrape the pages of the ELI Database of State IAQ Laws pdf document to create Dataset 2, Number of State Regulations. This dataset contains a count of how many air pollution regulation laws each state has in place. We used this dataset against our Dataset 1 which included data on chronic illnesses.

[2] ELI's Database of State Indoor Air Quality Laws: Main Page



ENVIRONMENTAL LAW INSTITUTE
DATABASE OF STATE INDOOR AIR QUALITY LAWS

COMPLETE DATABASE

Environmental Law Institute®
Copyright © 2022
Washington, D.C.

March 2022

Image of the ELI Database of State IAQ Laws Dataset PDF

The CDC: Annual State-Level U.S. Chronic Disease Indicators dataset used in our study was provided by the Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, and the Division of Population Health. The chronic disease indicators are surveillance indicators that were developed by consensus among the CDC, the Council of State and Territorial Territorial Epidemiologists (CSTE), and the National Association of Chronic Disease Directors (NACDD). The data comes from the various locations where a person may receive healthcare, i.e. hospitals, laboratories, doctors' offices, etc., and each jurisdiction that the CDC collects data from creates their own data sharing agreements with CDC. Therefore, each city, county and state may decide what information is collected and how/when it is shared with CDC.[3]

The extent to which participants are aware of their data collection/use may depend on individual city/county/state regulations. Since this data is collected in this fashion, it is also difficult to determine if any groups were systematically excluded.

When thinking carefully about the unit of analysis in the dataset we do not have data at the level of a single person. For our analysis, we explored other options including census tracts, county, and state. Each row represents a census tract which we will use as our unit. The granularity of data is based on years and location. Since our granularity is at an yearly level, our findings may not be as precise as they could be with data that is collected at a daily level.

Additionally, since we are focusing on the state level for our research questions, our findings may be less accurate for city or county level applications of our questions. When interpreting our findings, we must take these factors into account.

An unavailable feature we would like to have is the percentage of PM2.5 concentration in air pollution for each state. It would help us evaluate how PM2.5 is related to COPD by considering PM2.5's relation to air pollution in that location as a whole. We also wish we had PM2.5 concentration for state over time to simulate real time changes in PM2.5.

[3] https://www.cdc.gov/surveillance/projects/dmi-initiative/where_does_our_data_come_from.html

2. EDA

In order to get a better understanding of our data, we conducted exploratory data analysis (EDA). We visualized two quantitative variables (PM2.5 emission(annual mean), average rates of COPD) and two categorical variables (state, hospitalization reason).

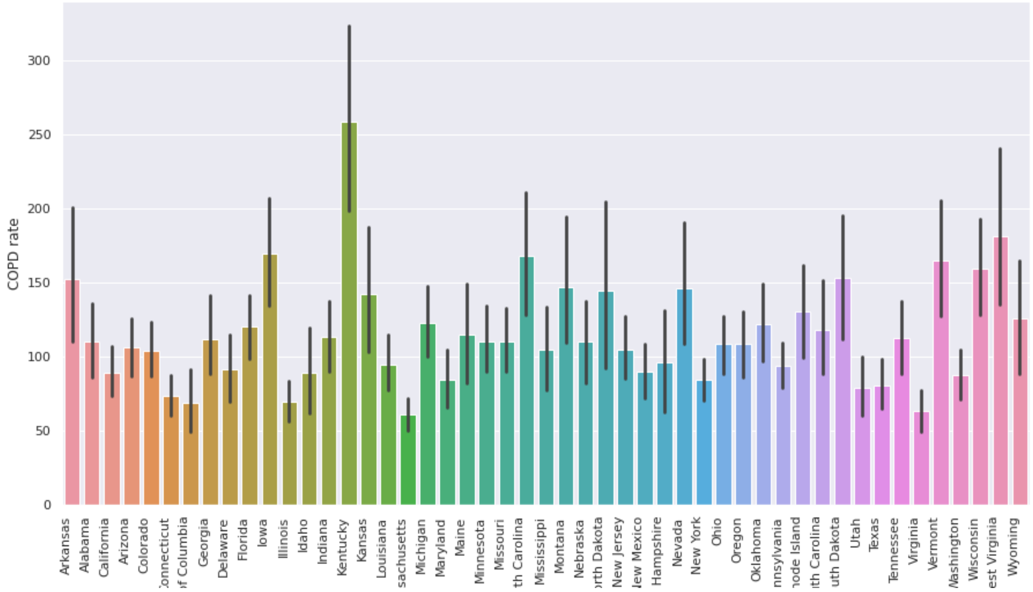


Figure 3. Average COPD rates based on state.

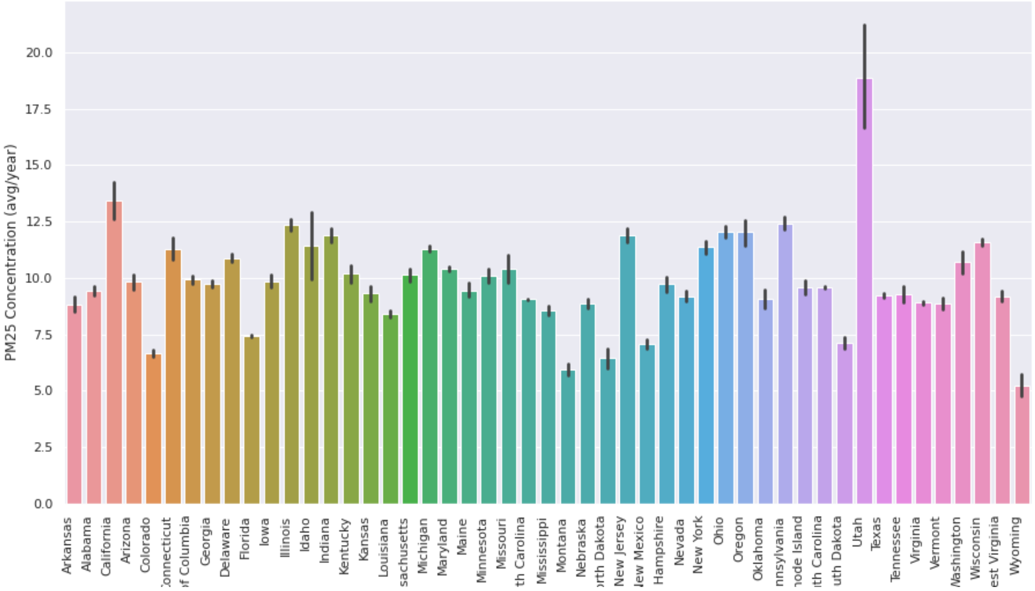


Figure 4. Yearly PM2.5 concentration levels based on state.

We may want to follow up on average COPD rates' and yearly average PM2.5 concentration levels' confounding factors.

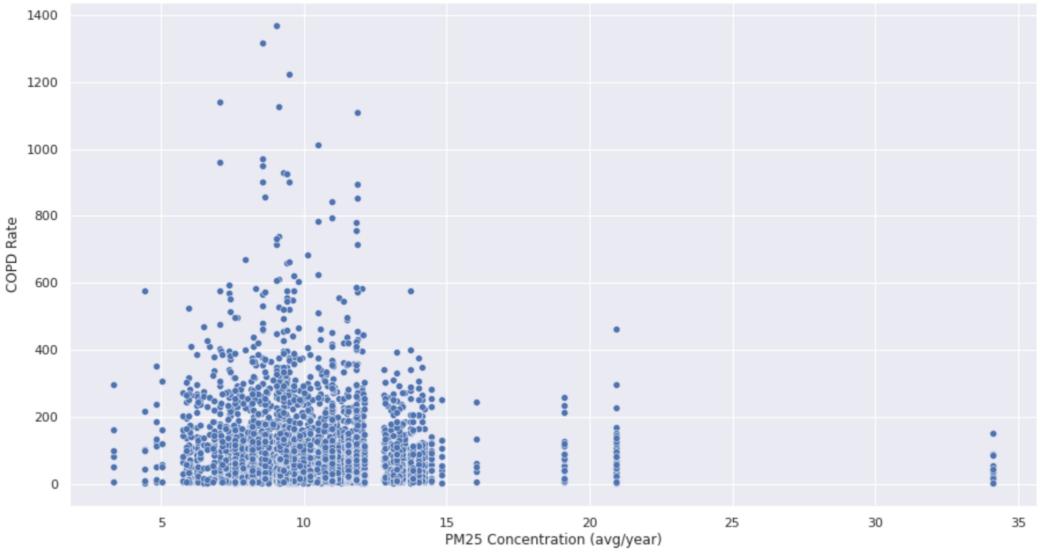


Figure 5. Yearly Average PM2.5 Concentration levels vs COPD rates

As seen in Figure 5, the relationship between yearly avg PM2.5 levels and COPD shows majority of data points clustered in the lower left quadrant of the scatter plot: lower COPD rates have lower yearly average PM2.5 concentration levels.

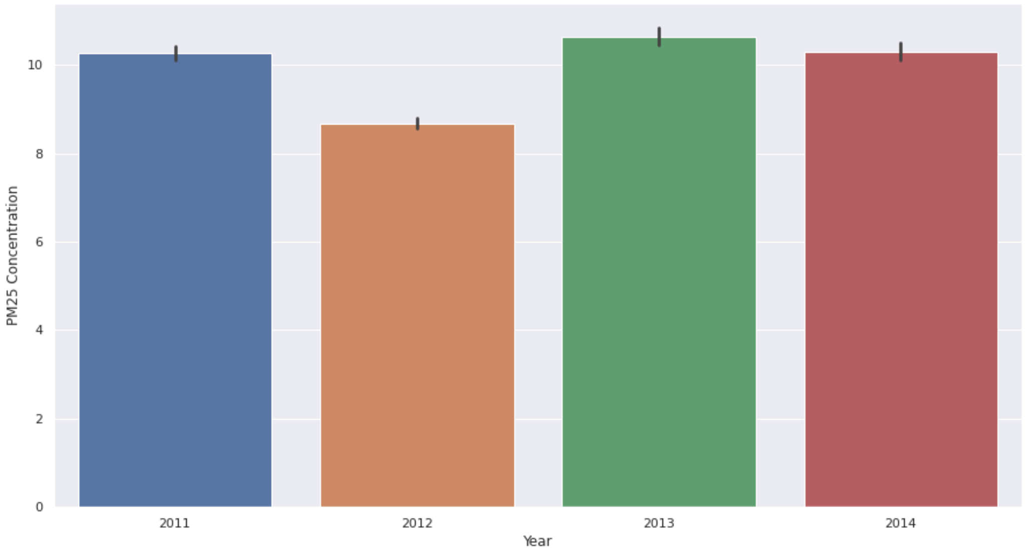


Figure 6. Average PM2.5 levels per year

We also saw that the yearly average PM2.5 concentration levels are not largely different across years, as seen in Figure 6. It appears the years 2011 and 2014 have similar average PM2.5 concentration levels. And the year 2012 has the lowest average PM2.5 concentration level whereas 2013 has the highest.

Kentucky appears to have the highest average COPD rate out of all the states as seen in Figure 3 and Utah appears to have the highest yearly average PM2.5 concentration levels out of all the states as seen in Figure 4.

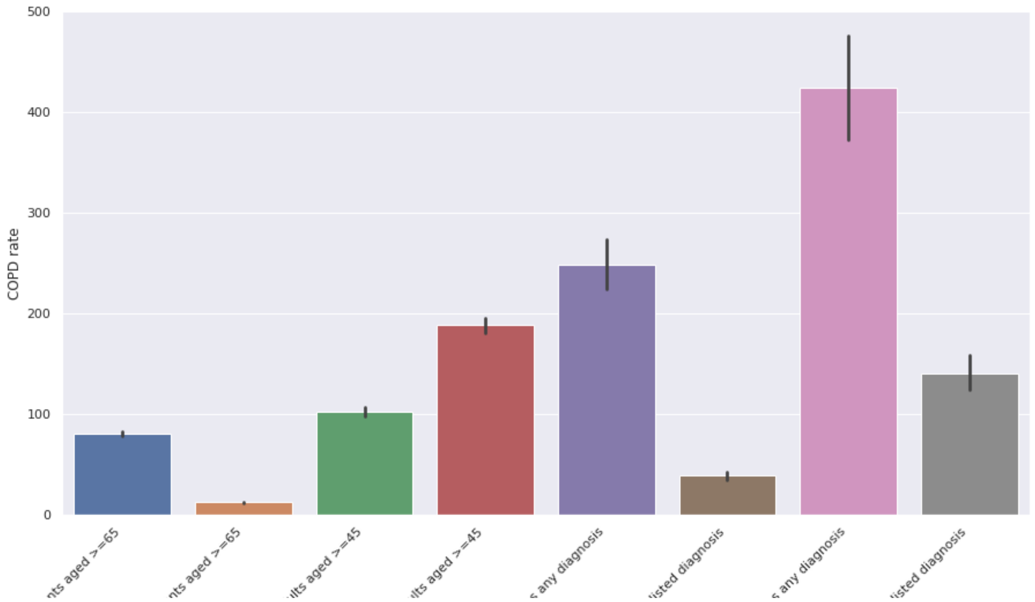


Figure 7. COPD rates based on hospitalization reasons

Additionally, we noticed that the hospitalization reason 'Emergency department visit rate for chronic obstructive pulmonary disease as any diagnosis' stands out as having the highest COPD rate when referring to Figure 7. The hospitalization reasons 'Hospitalization for chronic obstructive pulmonary disease as first-listed diagnosis among Medicare-eligible persons aged >= 65 years' and 'Hospitalization for chronic obstructive pulmonary disease as first-listed diagnosis' have significantly lower COPD rates.

When loading the CDC COPD rates and PM25 Concentration datasets, there were multiple columns with only null/nan values, so the first step in cleaning the datasets was removing these columns. For the CDC COPD dataset, we looked at the Data Value Type that was specifically Age-adjusted Rate and created datasets that looked at COPD rates based on Race/Ethnicity and Gender, which may be confounders. For the PM25 concentration dataset, we loaded the Daily Census Tract-Level PM2.5 Concentrations for each year between 2011 and 2014 separately due to size. For each individual dataset, we grouped by state FIPS and took the mean over ds_pm_pred, which is the mean estimated 24-hour average PM2.5 concentration in µg/m3. We then merged the PM25 concentration dataset and COPD rates dataset based on year and state FIPS (PM25 concentration DS) or LocationID (COPD rates DS). A decision made when cleaning the data that will impact our model and decisions was taking the annual mean by state PM2.5 emission; however, aggregating smooths out the data and allows for a more accurate analysis. We also converted all of the COPD rates into percentages in order to have a uniform dataset. To do this, we converted rates given in cases per 100,00 people, cases per 10,000 people, and cases per 1,000 people into percentages to avoid comparing data given in different formats.

Research Q1: Do increased levels of air pollution correspond with increased levels of Chronic Obstructive Pulmonary Disease (COPD)?
We compute the average difference in rates of COPD in high air pollution areas and compare it to the average difference in rates of COPD in low air pollution areas. The Figure 5 is a simple graphic that shows trends relevant to levels of COPD. We can also look into splitting up the data by year to see how COPD levels change over time if at all. From this graphic, we can see that a simple comparison of PM2.5 levels to COPD rates has no clear correlation. This is likely due to confounding factors like location, age, and pre-existing conditions.

Research Q2: Do increased regulations correspond with decreased chronic illnesses?
The unit we are using to look at regulations is at the state level. Figures 3 and 4 looking at the average COPD and PM2.5 concentration levels are separated by state. They will give us a basis to look at state regulations. State regulations is an instrumental variable because regulations can have an effect on air pollution levels but pollution regulations will have no direct impact on the rates of COPD.

3. Research Questions

Research Q1: Do increased levels of air pollution correspond with increased levels of Chronic Obstructive Pulmonary Disease (COPD)?[4]
We can use our research question to answer real world decisions regarding monitoring air quality in real time to develop technology for air cleaning products that will protect the environment from pollutants that potentially influence harmful disease prevalence.

Research Q2: Do increased regulations correspond with decreased chronic illnesses?[5]
A real world decision relevant to our research question involves regulations to place at the county and statewide level.

Research Q1 Approach: Causal inference
Better answered using causal inference as compared to multiple hypothesis testing / decision making as we originally explored. We can compute the average difference in rates of COPD in high air pollution areas and compare it to the average difference in rates of COPD in low air pollution areas.

Research Q2 Approach: Multiple hypothesis testing / decision making
Methods for dealing with multiple testing will call for adjusting α so that the probability of observing at least one significant result due to chance remains below our desired significance level.

4. Inference and Decisions

Research Q1 Approach: Causal Inference

Methods

Treatment and outcome.
Our Control Group consists of Low air pollution values and our Treatment Group consists of High air pollution values. In the article “ Understanding global PM2.5 concentrations and their

[4] Understanding global PM2.5 concentrations and their drivers in recent decades (1998-2016) - ScienceDirect
[5] Data:ELI's Database of State Indoor Air Quality Laws: Main Page

drivers in recent decades (1998–2016)," 10 ug/m^3 is listed as the threshold for "low" air pollution. The variables that correspond to the outcome of this test are the rates of COPD.

Confounders.
The confounders in our dataset include location, year, race/ethnicity, and hospitalization reasoning. The unconfoundedness assumption means we observe all the relevant confounding variables. I.e. there are no unobserved confounders. The unconfoundedness assumption does not hold. Intuitively and based on our results there should be other variables that have an effect on both treatment and outcome.

Adjustment for confounders.
One technique introduced in lecture used to adjust for confounders is matching. Each observation will need to have an exact match otherwise the data points will be skewed. However, we could not figure out how to use the matching algorithm with this dataset. To combat our confounders, we used a linear model instead:

COPD Rate = tau + a *LocationID + b*Race/Ethnicity + c*Hospitalization Reason +d*Year

This accounted for our prediction of the 2 main confounders: Race and Location. However, after further analysis, a linear regression approach and inverse propensity score weighting were better techniques for adjusting for our confounders.

Colliders.
In our dataset location and race/ethnicity are independent causes of hospitalization—the collider (since the two arrowheads collide into hospitalization).

Results

Interpretation.
We saw that linear regression will not produce an accurate prediction because we got small R^2 values and a very negative log-likelihood. After using logistic regression and inverse propensity weighting, we came up with an ATE value of -0.75. This result indicates very low causality between PM2.5 concentration and COPD rates. We used yearly state averages of COPD rates, which may have decreased the accuracy of our regression due to over-smoothing of the data.

Evidence against the hypotheses.
Our hypothesis was disproved by the investigation we did. Our ATE value of -0.75 indicates that increased levels of air pollution does not correspond to increased rates of COPD. We find this result surprising and we think there are other confounders that we did not account for that are adjusting our results. Without a more robust dataset, we struggled to find a result consistent with our common-sense expectation: that air pollution increases COPD rates. One explanation for this result could be that increased air pollution leads to higher rates of COPD, but PM2.5 is not the specific pollutant that affects rates of COPD.

Discussion

Limitations.
One of the limitations of our analysis is that we did not control for gender, which is a confounder. One of the limitations of matching is that for each observation we need to have an exact match otherwise the data would be skewed. Since our dataset is very complex, we could not use this approach. Therefore, we were limited to a linear model approach, logistic regression, and inverse propensity weighting.

Additional data.
It may be useful to include an additional column that accounts for gender when answering our causal question because it may be a confounder that has a significant impact on the outcome. Additionally, we could look at county level data as opposed to state level data to improve our accuracy.

Causal relationship between chosen treatment and outcome.
We are not confident about our causal relationship because of our strange ATE value. It is likely that there are more confounding variables that we did not consider that are affecting our predictions because our ATE value does not line up at all with our hypothesis.

Research Q2: Multiple hypothesis testing / decision making

Methods

Hypothesis.
The hypothesis we are testing using our dataset is "increased regulations correspond with decreased rates of chronic illnesses". We are asking the same question for 6 different chronic illnesses: Cardiovascular Disease, Asthma, Chronic Obstructive Pulmonary Disease, Cancer, Tobacco, and Alcohol, and we are looking at regulations at a state level. It makes sense to use multiple hypothesis testing because the amount of regulations may have a different effect on different diseases. Some diseases may be directly related to the pollutants that the regulations are managing, so we may see that increased regulations cause some diseases to decrease. However, we may also observe that some diseases do not have a significant correlation with pollutants that are being regulated. Therefore, it is important to consider multiple hypotheses where each chronic illness we are analyzing is tested individually.

Hypothesis test.
We will be testing each hypothesis the same way, using T-test. Since we do not have a binary outcome, such as a success/fail outcome, we cannot use methods like A/B testing. We have a distribution of outcomes, therefore T-test is the best approach. Our null hypothesis is that the difference in means equals 0, meaning our control and treatment group have equal averages. Our test statistic is $control\ group\ mean\ -\ treatment\ group\ mean$. We anticipate a positive number for the test statistic across all of our hypotheses since higher rates of regulations may correspond to lower prevalence of chronic illness. Therefore, we expect the control group mean to be larger than the treatment group mean.

To partition the dataset into control and treatment groups, we graphed the number of regulations by state, as seen in the figure below, and concluded that a threshold of 5 regulations gave us a roughly even distribution of control and treatment groups: there are 22 states with 5 or more regulations in the treatment group and 29 states with less than 5 regulations in the control group. We saved the states in the control and treatment groups in two separate dictionaries, which we referred to when partitioning the data by chronic illness. Our dataset of chronic illnesses contained significant data on the raw number of cases, so we only considered that data value type. To test each hypothesis, we created a new dataset from the cleaned CDC chronic illnesses dataset which only highlights one disease of interest. This new dataset is then divided into a control and treatment group by state: if the state abbreviation is in the control state abbreviation dictionary, then it is assigned to the control table for that particular illness, and the same method was used for the treatment group.

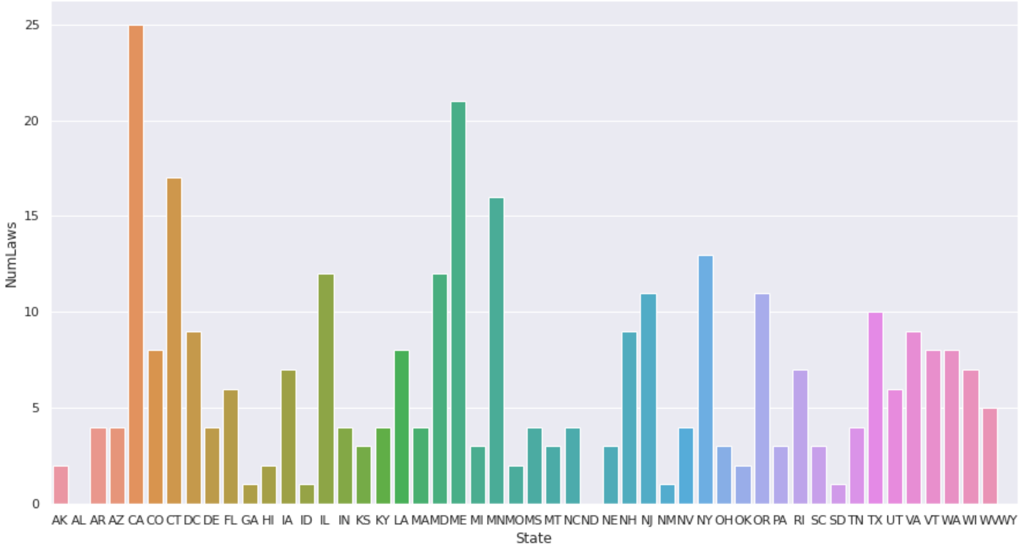


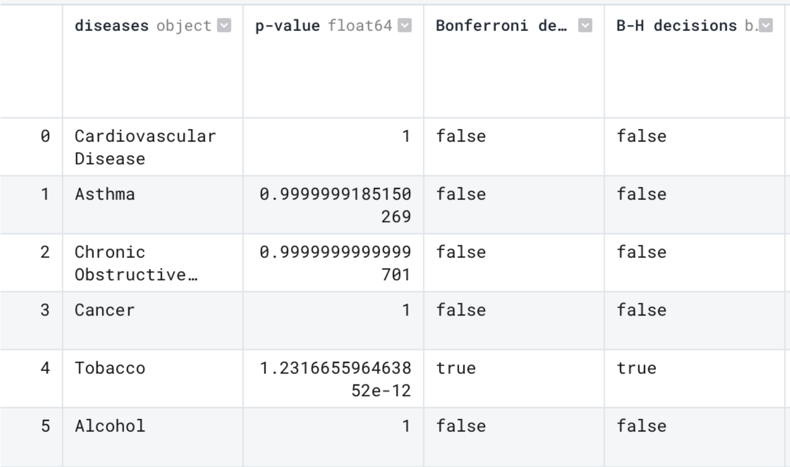
Figure 8. Distribution of regulations by state.

Correct for multiple hypothesis tests. Error rates being controlled.
We will be correcting for multiple hypothesis tests using the Benjamin-Hochberg and Bonferroni Correction procedures. The Benjamin-Hochberg method controls for the false discovery rate (FDR) among our 6 hypotheses: our desired $FDR \ = E[\# \, false \, positives \ / \ \# \, positives]$. For testing our 6 hypotheses with FDR less than or equal to our α, 0.05, we declare a discovery for all p-values with value less than or equal to our calculated threshold. The Bonferroni method controls for the family-wise error rate (FWER) among our 6 hypotheses: our desired $FWER \ = \ P(at \, least \, one \, false \, discovery)$. For testing our 6 hypotheses with FWER less than or equal to our α, 0.05, we test each hypothesis with significance $\frac{0.5}{6}$.

Results

After conducting our tests with naive thresholding, we accepted the null hypothesis for 5 out of 6 of our tests since the p-value was significantly bigger than 0.05. Therefore, it is almost certain that the amount of regulations does not have an effect on the rate of chronic illnesses. The one chronic illness for which we rejected the null hypothesis was tobacco use. The decisions we came up with for both Benjamin-Hochberg and Bonferroni procedures both follow the same results as the naive threshold results. The results can be seen in Figure 9 below. We notice that the number of regulations does appear to affect the rate of tobacco related chronic illness but it has no effect on the rate of other diseases. A possible confounder is that tobacco is a highly regulated substance in the United States whereas other chronic diseases are not. Additional confounders include what the regulations say, when they were enforced, and how they are enforced.

In our results, we have negative test statistics corresponding with really high p-values. The conclusion we came to is that average rates of chronic diseases in our control group are lower than the average number of diseases in our treatment group; recall that our treatment group corresponds to states with a high number of regulations. Overall, we fail to reject the null hypotheses; therefore we reject the alternative hypothesis: higher rates of regulations may correspond to lower prevalence of chronic illness. However, we are not saying that low numbers of regulations correspond to low numbers of diseases since correlation is not causation. Therefore, the number of regulations does not have an effect on the number of chronic illnesses. What we are likely observing is that there is no relationship between regulations and chronic illnesses.



	diseases object	p-value float64	Bonferroni de…	B-H decisions b…
0	Cardiovascular Disease	1	false	false
1	Asthma	0.9999999185150269	false	false
2	Chronic Obstructive…	0.9999999999999701	false	false
3	Cancer	1	false	false
4	Tobacco	1.231665596463852e-12	true	true
5	Alcohol	1	false	false

Figure 9. Results from Multiple Hypothesis Testing

Discussion

Our p-values would be different if we did not specify an alternative hypothesis to align with our research question: we are looking for an alternative hypothesis that says increased numbers of regulations corresponds to decreased number of illnesses. This means our control group must have a greater average. Otherwise, our alternative would just be that a difference exists between the control and treatment group.

Significant discoveries.

After applying our correction procedure, tobacco remained significant in both cases.

Limitations.

Since we are doing this analysis at a state level, there will be regulations at lower levels we are not considering. We are looking at the number of regulations, not the content of those regulations. Perhaps one regulation is really effective but many others are not. We are not able to quantify the effect that the regulations are having since it is too subjective and there are too many confounders to be able to assign a numerical value to the effectiveness of a regulation.

Additional tests.

With additional data we would perform multiple hypothesis testing at city or county level instead of state level; the results would be more accurate since we would have an increased number of data points. If we were able to quantify the effects of each regulation, we would also be able to get a more accurate idea of how government regulation can be used to reduce rates of chronic diseases.

Conclusion

Key findings.
When using causal inference to answer our question "do increased levels of air pollution correspond with increased levels of Chronic Obstructive Pulmonary Disease (COPD?", we acquired an unexpected result. After using logistic regression and inverse propensity weighting, we obtained a negative ATE value which indicates very low causality between PM2.5 concentration and COPD rates. We may be getting this unexpected result because COPD is not caused by PM2.5. COPD rates likely increase with increased air pollution as a whole and PM2.5 is just one component of air pollution. Therefore, there is essentially no correlation between COPD and PM2.5, which is evident in our negative ATE. However, that does not mean that there is no correlation between air pollution as a whole and COPD rates.

Additionally, when applying multiple hypothesis testing to answer our second question "do increased regulations correspond with decreased chronic illnesses?", we were surprised by our results. We fail to reject the null for 5 out of 6 of our hypotheses. In the case of the chronic illness corresponding to tobacco use, we rejected the null hypothesis even when correcting for error rates through Bonferroni and Benjamin-Hochberg procedures: the number of regulations may impact the prevalence of the chronic illness associated with tobacco. However, we are likely witnessing the effects of confounders, especially considering that tobacco is a highly regulated substance while other chronic illnesses are not regulated in the same fashion. Overall, our results indicate that there is no correlation between regulations and prevalence of chronic illnesses.

Our results are generalizable and reproducible at a state level since we are only considering data at a state level. At anything smaller, such as at a city or county level, it is not generalizable or reproducible since we are not considering smaller units.

Our findings are broad since both of our hypotheses were proved incorrect. There is not a positive correlation between PM2.5 and rates of COPD and there is not a negative correlation between increased regulations and chronic illness rate. We are not able to make correlation statements because we disproved our hypothesis for both research questions.

Call to action.
Based on our results, we believe air quality and chronic disease trends should be used to inform health regulations and air cleaning technology creation and access. Decisions and actions that can be taken are monitoring air quality in real time to develop technology for air cleaning products that will protect the environment from pollutants that potentially influence harmful disease prevalence.

Additional data sources benefits.
The CDC: Daily Census-Tract PM2.5 Concentrations dataset was too large to combine. The PM2.5 concentration data was broken down by year when downloaded so we used yearly average to combine into one dataset for our analysis. The benefit is that it worked for us since we were using yearly average in answering our research questions. The consequences is that total yearly data was not included which would only be a problem if our research questions were using a different test statistic.

Limitations.
Limitations were in part the available units and granularity of the data. At the beginning of our analysis, we wanted to explore at individual levels (such as for an individual person) the PM2.5 concentrations and COPDS rates to understand chronic illness prevalence and preventative measures. As the level of individual data was not available, we exlcuded it from our analysis. Perhaps it could be incorporated when exploring these same research questions in a smaller more local dataset such as Berkeley and comparing it to neighboring cities like Albany.

Future work.

Future studies that could build on our work are studies that suggest that chronic diseases may be aggravated by air pollution. For example, with the fire that occurred at the Chevron refinery in Richmond CA a few years prior we could explore the increased levels of air pollution as a result of such events by looking at the air pollution trends over time to see aggravation of COPD and compare it to another disease like asthma.