Data 102 Final Project

Group 1: Alex Long, Cat Tseng, Tanya Saharan, Trigger Nandhra

Data Overview

For our research project, we utilized the CDC's 2023 Release of the U.S. Chronic Disease Indicators (CDI) dataset, derived from indicators outlined in the Morbidity and Mortality Weekly Report (MMWR). This dataset amalgamates data from various sources, but the mortality data for cardiovascular disease (CVD) and chronic obstructive pulmonary disease (COPD) used in our project is sourced from the National Vital Statistics System (NVSS), comprising over 100 million state death records contained in the National Death Index (NDI).

Given that the CDI dataset encompasses data from all governing state bodies within the U.S., it constitutes a census. During our exploratory data analysis of cardiovascular disease mortality from the CDI dataset, we observed racial exclusion. Notably, mortality data for Asians were unavailable in 16 states, Asian Pacific Islanders in 9 states, African Americans in 7 states, and Hispanics in 6 states. No explanation was provided for these data gaps. Since Asian Americans exhibit lower mortality rates from heart disease compared to white adults (US OMH 2021), this disparity could potentially bias our causal inference analysis for cardiovascular disease mortality, particularly considering the significant representation of white adults in the racial demographic of Southern states.

Moreover, the state government's census data may overlook certain segments of the population, including undocumented individuals like illegal immigrants, and those lacking birth/death certificates. Only communities with higher socioeconomic status can typically afford medical care for chronic illnesses, given the prohibitive costs of healthcare in the U.S. for

marginalized groups. Hence, the data could disproportionately represent wealthier or more politically visible demographics.

Regarding the CDI dataset, it is uncertain whether each participant was informed about the collection and use of their data, as the CDC's data seems to be aggregated from various sources. The NDI data originates from state birth and death records, indicating some awareness among participants regarding their data being in a government database.

The granularity of the CDI is limited, with each row representing an abstracted statistic concerning a particular chronic disease indicator ("Topic" and "Question") among a distinct demographic ("StratificationID1") in a given location and time ("LocationAbbr" and "YearStart"). Documentation explaining the column variables was lacking, requiring us to infer their purpose from contextual clues. For example, "DataValueType" includes ambiguous values, necessitating selective filtering of rows with unclear data.

For Question 2, we focused solely on the "DataValueType" labeled as "Number", presuming it represents the raw count of individuals categorized within a specific chronic disease indicator. Within the "Question" column, various mortality categories fall under the "Cardiovascular Disease" topic, so we only selected statistics from "Mortality from total cardiovascular diseases" to avoid double counting. These considerations will add to the uncertainty of our results, as we will need to draw conclusions based on limited specificity.

Since the data collection methods of the "DataValue" column were undisclosed, we cannot definitively assess the presence of selection bias, measurement error, or convenience sampling. Convenience sampling may exist, given that death record data is mostly collected within the healthcare system. Since we assume that "DataValue" denotes the number of

individuals who died from the specified chronic disease (cardiovascular disease or COPD), measurement errors seem inevitable for such a considerable statistic.

While it remains unclear whether the CDI dataset undergoes modification for differential privacy, its abstracted nature inherently safeguards the privacy of included individuals. Data values were devoid of specifics about any particular individual. Moreover, the dataset's low granularity precludes the extraction of identifiable information through data analysis, offering only state-level locations, annual summary mortality statistics, and broad stratifications such as gender and race/ethnicity. Absent more detailed information, no personal or private data can be discerned.

The CDI dataset lacks data concerning age, genetic predisposition to COPD and CVD, and some demographics, all influential variables for our analysis, as they would enable the isolation of our research variables from confounding factors.

Before conducting analysis, we filtered our data on the starting year, chronic illness, and type of data value using the columns "YearStart", "Question", and "DataValueType." The selection of "YearStart" over "YearEnd" was arbitrary. Within the "Question" column, we pinpointed the particular disease (e.g., "Mortality with chronic obstructive pulmonary disease as underlying cause among adults aged >= 45 years" for COPD in our initial question), ensuring the data pertained to deaths. We used the "Number" data type to avoid interpreting the ambiguous "DataValueUnit" column.

In our first question, we cleaned our data by filtering the "Location" column to cross-reference states in both the CDI and EPA datasets. We grouped the data by state and year, aggregating by the mean to exhibit the average number of adult mortalities attributed to COPD for each stratification. For the second question, we designated treatment values ('0' or '1') to

classify states as "Southern," as defined by the U.S. government. Additionally, we filtered the "StratificationCategoryID1" column for "OVERALL" values to avoid duplicate data.

In addition to the provided CDI data, we chose to add air quality data from an extraneous dataset compiled by the EPA (Environmental Protection Agency). This dataset tracked the presence of four major pollutants (NO2, O3, SO2, and CO in parts per billion) across certain cities in the United States four times a day, every day from 2000 to 2016. We wanted to cross-reference the EPA's data with the CDI data and examine the relationship between the concentration of pollutants and the mortality with chronic obstructive pulmonary disease as an underlying cause among adults (aged >= 45 years) in certain U.S. states.

The data was collected from samples of a certain location's air quality on a specific day. Viewing the distribution for the concentration of O3 in Phoenix, Arizona, we observe an upward trend. However, we could not find data regarding the concentration of O3 across the entire state of Arizona. Therefore, we expect that the EPA data may not have high generalizability, as the location, date, or time the data was collected may not accurately represent the entirety of the state or even Phoenix.

Since the EPA's dataset did not consist of any demographic information, we can conclude that no groups were systematically excluded from the data. Furthermore, differential privacy cannot be applied because there are no individual participants in the population to be aware of. However, for the participants collecting the data, we assume that they are aware their findings are being used for research purposes since their compilation is publicly available.

The granularity for the EPI's data is much higher than in the CDI dataset. Each row represents multiple statistics aggregated from the amount of NO2, O3, SO2, and CO in parts per billion in a particular location in a U.S. state during a specific day and time. This will impact the

interpretation of our findings, as the specificity of this data makes it much easier to analyze and cross-reference in correspondence with the CDI's data.

It is almost certain that the equipment the EPA used to measure air quality produced inaccurate readings at some point, leading to measurement errors. Moreover, each location was likely strategically placed for ease of access or maintenance, introducing a source of convenience sampling. Therefore, measurement error and convenience sampling are relevant in the context of the EPA dataset.

Features/columns regarding the cardinal regions where the EPA data collection took place would help answer questions such as which region of the U.S. has the highest concentration of specific pollutants. There were no columns with missing data in this dataset.

During data cleaning, we first extracted rows based on relevant years that appear in both datasets. Next, we isolated data from the top 25 most urban states as defined in "America's Most Urban States". We subsequently grouped our data by relevant states and years and aggregated by the mean of the average concentration of each pollutant. Finally, we merged the modified CDI and EPA datasets on state and year to produce a single data table. While preprocessing data is practical for analysis, it can oversimplify the complex trends within our models and inferences, as aggregating on already abstracted data can result in incorrect predictions and conclusions.

EDA

RQ1 Qualitative Visualization:

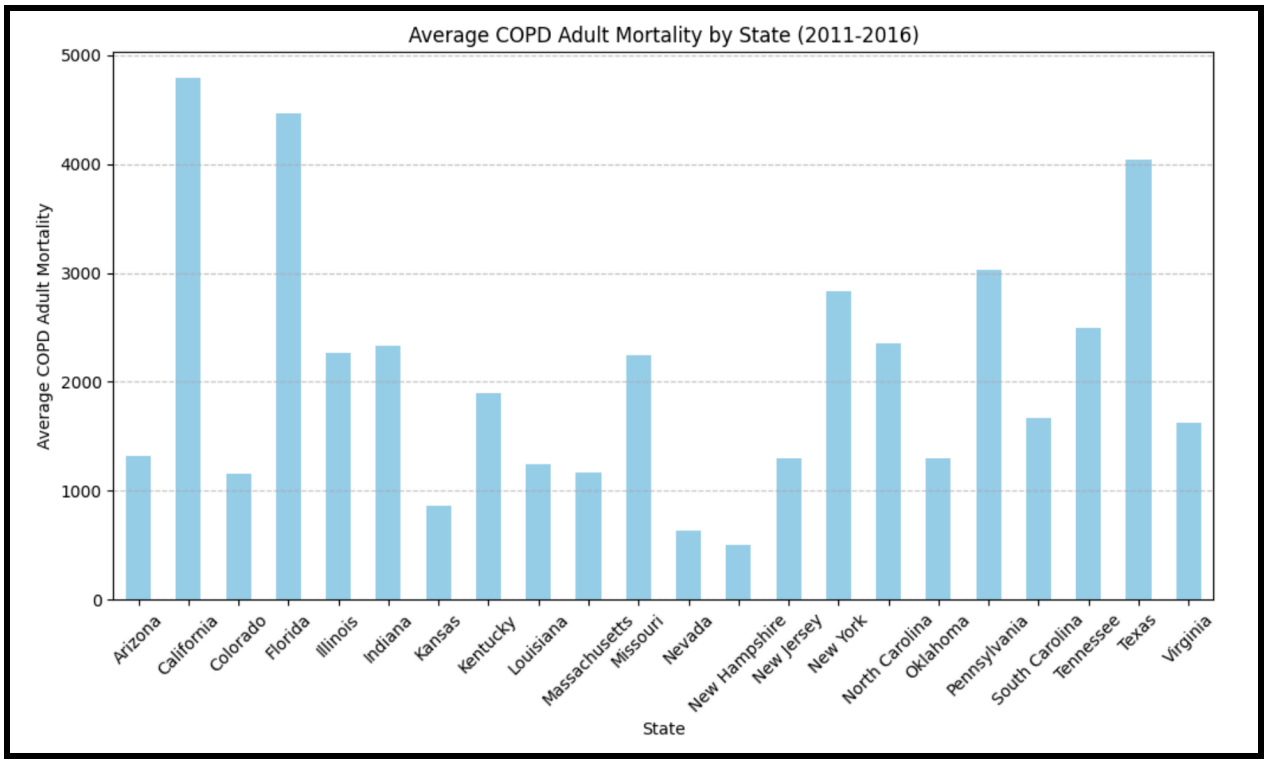
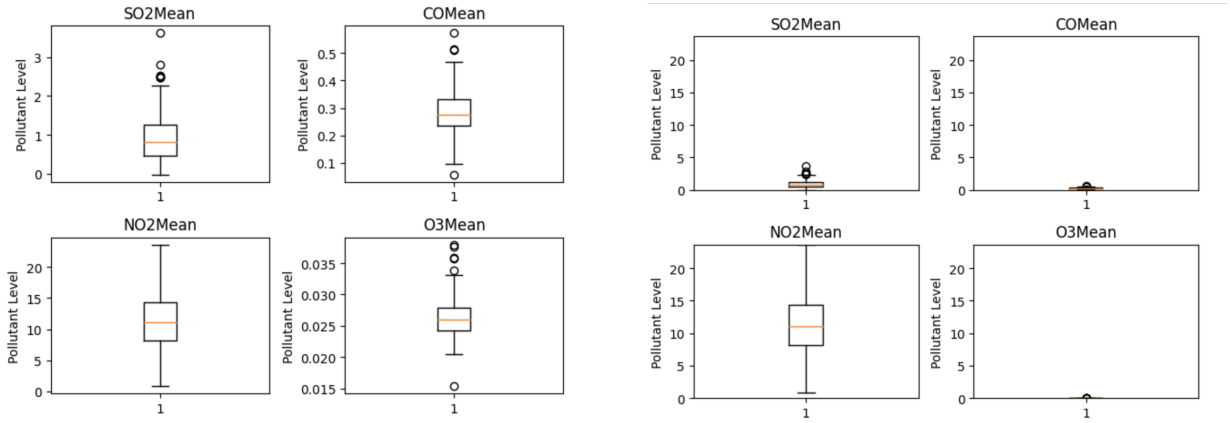


Figure 1**: **Average Adult Mortality Rates from COPD
Bar plot displaying average mortality rates in urbanized states from the years 2011-2016.

We see that California, Florida, and Texas have the highest average COPD mortality. We question whether this is because of their large populations or more comprehensive data collection methods. We could also investigate whether some states have greater deaths from COPD by examining the quantity of each state's data.

Our research question asks how the frequency of certain pollutants affects the number of mortalities caused by COPD among adults in more urbanized states. Before we begin the analyses, knowing which states have the highest mortalities, allows us to understand the context of our data better.

RQ1: Quantitative Visualization:



Different y axes (left). Same y axes (right).

Figure 2a (left), 2b (right)**: **Distribution of 4 pollutants across urbanized states
Box plots displaying the spread of pollutants Sulfur Dioxide (SO2), Carbon Monoxide (CO), Nitrogen Dioxide (NO2), and Ozone (O3) across urbanized states from years 2011-2016. Figure 2a has different y axes. Figure 2b has a standardized y-axis.

First, we analyze the spread of each pollutant. Sulfur dioxide is slightly skewed right. The rest are approximately normal. When the y axes are the same, it is clear that nitrogen dioxide has the highest levels, followed by sulfur dioxide, carbon monoxide, and ozone. This leads to the question of why nitrogen dioxide has the highest average and whether these averages look the same for each state.

This is another essential visualization that informs us about the data we're working with. It is important to properly understand the distribution of pollutants (which pollutant has the highest vs. lowest level of distribution) and how they compare to one another.

RQ2 Qualitative Visualization:

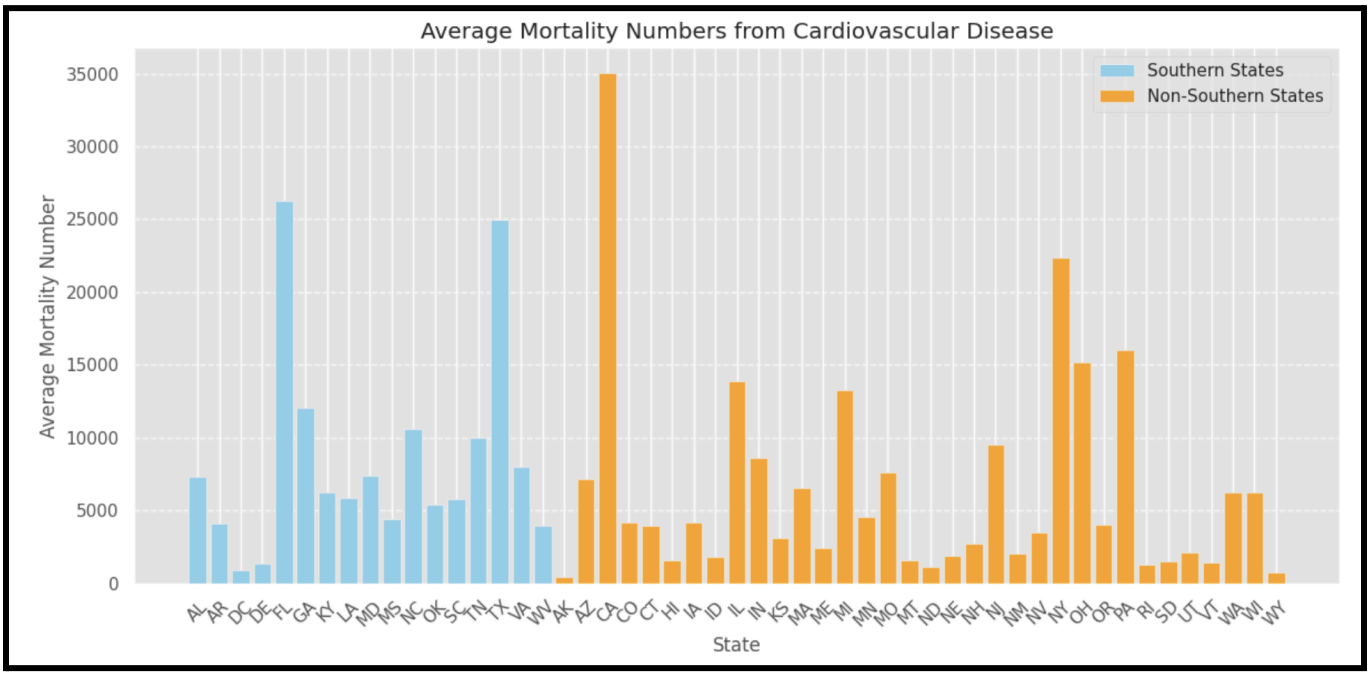


Figure 3**: **Average Mortality Numbers from Cardiovascular Disease in 2020
Bar plot showing average mortality numbers across all US states in the year 2020.

The Southern states have slightly higher averages than the other states when you exclude California from consideration. We want to see if the mean and median for Southern States are significantly higher than non-Southern states. Hypothetically, we could conduct a two-sample t-test to determine whether the underlying population difference is significant.

Our research question asks whether living in a Southern State has a causal effect on mortality from cardiovascular disease. We need this visualization to observe whether or not there are any visual differences in mortality rates between southern and non-southern states. From the graph above, there are not any notable differences.

RQ2 Quantitative Visualization:

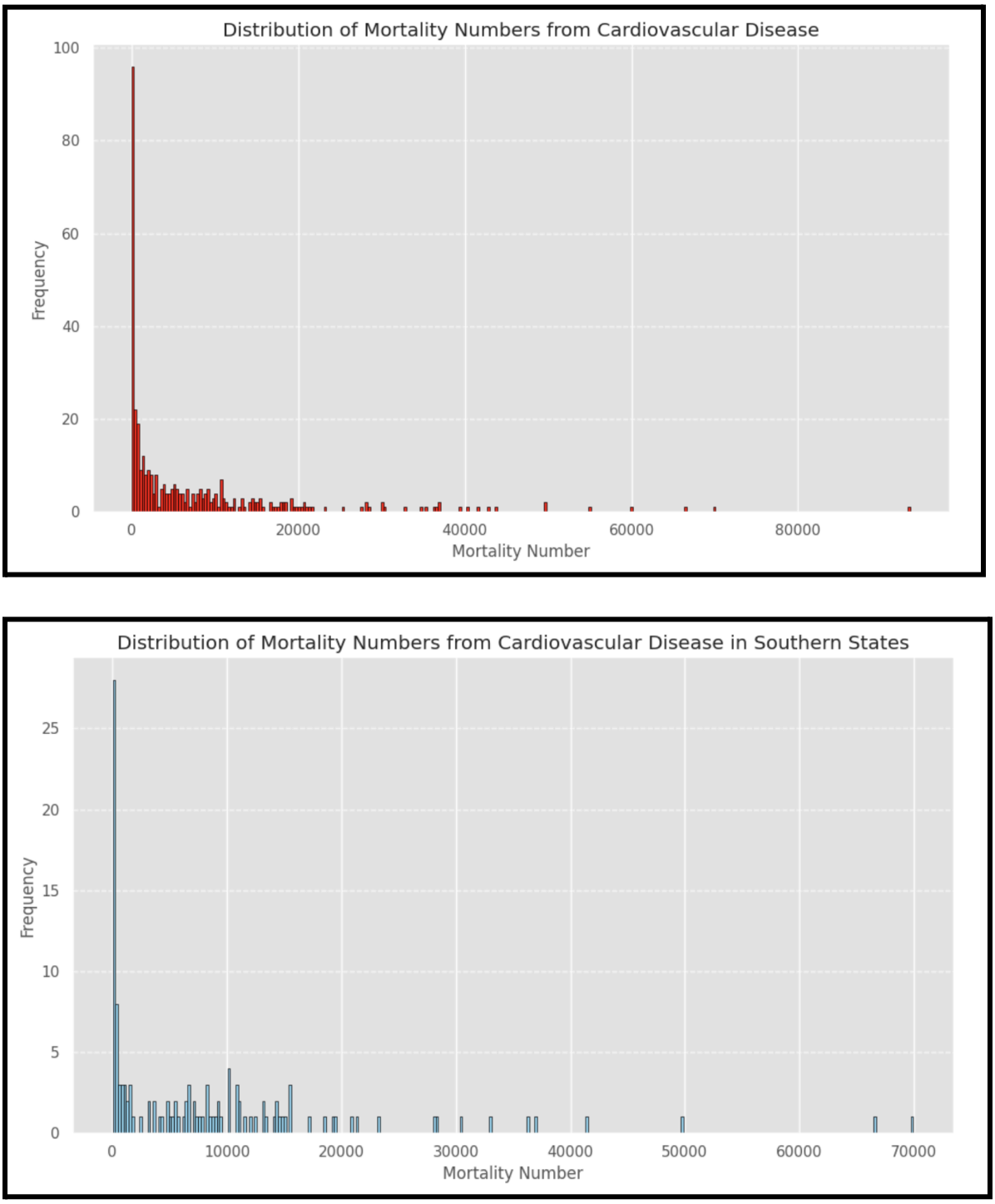


Figure 4a (top), 4b (bottom): Distribution of Mortality from Cardiovascular Disease
Histogram plotting the distribution of mortality numbers from CVD in all US states (fig 4a) and in only Southern states (fig 4b).

The highest frequency of mortality rates occurs below mortality counts of 500. Our first plot demonstrates outliers - the highest death count being well above 80,000. To identify factors contributing to higher deaths from CVD and COPD, we can explore which states have the outliers and when they typically occur. Our second plot explores this variance in only Southern States, which looks similar to the plot with all states; both are highly skewed rightwards.

This visualization enabled a better understanding of the variance in mortality from cardiovascular disease between Southern states and the entire U.S. After comparing both, there does not seem to be a difference, which hints that there may not be a causal effect from living in a Southern state on mortality from cardiovascular disease.

Research Question 1: Prediction with GLMs and non-parametric methods

For our first research question, we wanted to determine if we could utilize the frequency of certain pollutants (SO2, CO, NO2, and O3) to predict the number of mortalities with COPD as an underlying cause among adults (45 and older) living in more urbanized U.S. states. Referencing an external dataset extracted from the Census Bureau ("America's Most Urban States"), which ranked the U.S. states with the highest urban densities, we chose the top 25 states (New Jersey, California, Nevada, Massachusetts, Florida, Hawaii, Rhode Island, Utah, Arizona, New York, Colorado, Maryland, Illinois, Connecticut, Delaware, Washington, Texas, Oregon, Pennsylvania, Ohio, Michigan, Georgia, Virginia, Minnesota, Louisiana) for our analysis. Real-world decisions regarding federal resource allocation, environmental legislative action, and urban planning can be made by answering this question.

Methods

To generate the predictions, we utilized generalized linear models (GLMs). Since we can choose from several distribution families and link functions, we can compare and contrast multiple models to choose the best fit for our data. Furthermore, we can efficiently coordinate multiple features, allowing us to observe the performance and interaction of each on the outcome. Therefore, GLMs are an exceptional tool for investigating the complex relationship between pollutants and adult mortalities with COPD as an underlying cause. For our analysis, we used a GLM with a Negative Binomial distribution and a Log link function. We chose this model due to its practicality and flexibility when working with count data, which may or may not follow a normal distribution.

Although GLMs are a powerful tool, many limitations hinder their performance. For instance, as its name suggests, GLMs assume a linear relationship between the features and the link-function transformed expected outcome. Consequently, highly nonlinear relationships will dramatically impact the quality of our predictions. Moreover, there are limited distributions that a GLM can fit. Therefore, if the data provided are not somewhat distributed according to the above distributions, it can result in a poor fit. Lastly, GLMs assume that our features are independent, which means unexpected collinearity will skew our predictions. Therefore, despite GLMs' predictive power, egregious violations of their assumptions will result in a lackluster performance.

We are trying to predict mortality with COPD as an underlying cause among adults 45 years and older. The features we are using are the average concentration of pollutants SO2, CO, NO2M, and O3 in the air. We chose these features due to their scientific association with the increased risk of contracting COPD.

We chose to use a Negative Binomial distribution with a Log link function for our GLM, as the Negative Binomial GLM yielded a much smaller AIC (Akaike Information Criterion) than other relevant models, indicating higher predictive power. We can further justify our decision to use the Negative Binomial distribution, as our outcome (mortality with COPD as an underlying cause among adults aged >= 45 years), has a variation far greater than the mean. We assume linearity in our model parameters and independence among our GLM features. For our Bayesian GLM, we could not decide on a specific prior, so we chose to use the default prior distributions provided by Bambi.

We will use the random forests model as our nonparametric method, due to its high accuracy and management of nonlinear relationships. We assume that our features are relevant to our outcome and are not subject to high collinearity.

We will evaluate each model's performance by computing and comparing their mean absolute error (MAE) with respect to the test set.

Results

For our Negative Binomial GLM in a Frequentist perspective, we received p-values of 0.101, 0.290, 0.013, and 0.052 for our features SO2, CO, NO2, and O3 respectively. This insinuates that the effect of SO2, CO, and O3 on the variation of COPD adult mortalities is insignificant (>= 0.05), while that of NO2 is significant (<= 0.05). However, its negative coefficient (-0.0654) indicates a slight negative relationship between NO2 and COPD adult mortalities, which is unexpected. Moreover, our low pseudo-R-squared (0.0823) demonstrates that our features in general do not explain the variance of COPD adult mortalities very well, only accounting for 8%.

For our Negative Binomial GLM in a Bayesian perspective, we received averages of 1.495, -0.065, -58.395, and 0.246 for the mean concentration of CO, NO2, O3, and SO2 respectively. In addition, we received HDI intervals of (-0.190, 3.156), (-0.107, -0.024), (-114.24, -5.981), and (0.030, 0.461) for CO, NO2, O3, and SO2 respectively. From our results, we can deduce that CO and SO2 have a positive association with COPD mortalities, while NO2 and O3 have negative associations of varying strength with COPD mortalities. Moreover, the width of the HDI intervals also indicates substantial uncertainty with our predictions, especially with CO and O3.

For our Random Forest nonparametric method, we received an R-squared score of 0.373, meaning that 37.3% of the variance of our data can be explained by the model. Therefore, our model lacks significant predictive power.

We estimate that our models have substantial uncertainty in our GLM predictions. In our Frequentist and Bayesian GLMs, we computed a negative association between NO2 and COPD adult mortalities with reasonable certainty. However, for the rest of the pollutants, our results indicate substantial uncertainty. This variability is best shown through our Bayesian GLM analysis, as the HDI intervals produced by our model are noticeably wide for the average of the mean presence of SO2, CO, and O3. Therefore, for a majority of our features, we experience substantial uncertainty regarding our GLM predictions.

To summarize, our results indicate that CO, NO2, O3, and SO2 concentrations are unsatisfactory predictors for adult COPD mortalities.

Discussion

We compute MAEs of 1071, 1109, and 684 for our Frequentist GLM, Bayesian GLM, and Random Forest nonparametric models respectively. Therefore, our Random Forest model performed the best, displaying a noticeably smaller MAE. By design, the Random Forest model employs several decision trees and computes the average of their predictions. Although more complex, this nonparametric model is designed to be highly accurate, better capturing the complex nature of our data than our GLMs. We are confident in applying the Random Forest model to future datasets due to this model's high performance.

For our Frequentist GLM, we received a deviance of 48.5 and a Pearson chi-squared of 42.9. Using a deviance and Pearson chi-squared of 0 as our baseline for perfect predictive power, we conclude that this model does not fit the data well. Its low R-squared score of 0.082 is also

indicative of its poor fit. While our Bayesian GLM summary does not have a direct goodness-of-fit indicator, our earlier analysis implies that the model fits well for mean NO2 concentration, while fitting poorly for our other features. However, its large MAE implies a worse fit than our Frequentist model. Although our Random Forest model could only explain 30.7% of the variance of our outcome, it produced the lowest MAE of the three models, indicating a better fit than the Frequentist and Bayesian GLMs.

Research Question 2: Causal Inference

For our second research question, we explored the causal effect of living in a Southern state in the year 2020 on the number of people who died from chronic cardiovascular disease because many Southern states like Louisiana, Mississippi, and Alabama had the highest death rates due to total cardiovascular disease. If there is a causal effect, we could conduct further analysis into location-dependent factors that affect the incidence and severity of cardiovascular diseases and inform federal and state resource allocation (i.e. more healthcare resources in high-incidence regions or preventative medicine initiatives).

We chose causal inference to compare observed outcomes (cardiovascular disease mortality rates) with counterfactual scenarios (living in a Southern state or not). Causal inference is appropriate as it can provide impactful information on whether location directly contributes to cardiovascular disease, directing potentially lifesaving interventions. Most importantly, causal inference can help identify and mitigate biases in observational studies by accounting for potential confounding variables. This is particularly important in our question since randomized controlled trials are not feasible or ethical with the data available since the CDI dataset is observational and collected before our research project.

We used three methods to conduct causal inference - Propensity Score Matching, Inverse Propensity Weighting, and Inverse Propensity Weighting with Trimming 20% off the ends of the data. We also conducted a t-test with the null hypothesis stating that there's no difference in the mortality between Southern states and the rest of the US, and the alternative hypothesis stating that there is.

One limitation of using causal inference for our research question is that we cannot properly account for all of the confounding variables that may impact our analysis due to the

enormous amounts of missing data, especially for uncollected information (i.e. genetic predisposition). This failure to consider all relevant confounding variables can lead to biased estimates of causal effects. In addition, concluding group-level data can lead to erroneous causal inferences as the higher level of abstraction can mask individual variation and lead to incorrect conclusions. For example, the CDI dataset does not stratify into different socioeconomic groups, which prevents us from isolating the effect of living in Northern vs. Southern states from other factors like the distribution of socioeconomic classes across the U.S.

Methods

In this question, we will measure the causal effect of a binary treatment Z on an outcome Y by considering the potential outcomes Y(0) and Y(1). Treatment Z is living in a Southern state (1) or not living in a Southern state (0). Y(0) represents the total number of deaths from cardiovascular disease in non-Southern states. Y(1) represents the total number of deaths from cardiovascular disease in Southern states. The Southern states are Alabama, Arkansas, Delaware, the District of Columbia, Florida, Georgia, Kentucky, Louisiana, Maryland, Mississippi, North Carolina, Oklahoma, South Carolina, Tennessee, Texas, Virginia, and West Virginia.

Race and gender are confounders because they are both factors that affect genetic predisposition to cardiovascular disease and are correlated with socioeconomic status. Both can also impact lifestyle choices like diet and exercise. The unconfoundedness assumption does not hold because the dataset excludes some unobserved confounders. Confounders for mortality from cardiovascular disease could include genetic predisposition, culture/lifestyle (i.e. diet, exercise, SES, etc.), access to healthcare, and more. None of these are explicitly included in the dataset we are working with.

Below is a causal DAG for our research question. Since the intermediate factors are not included in the CDI dataset, the only relevant collider in our research question is the variable we are interested in predicting: mortality from total cardiovascular disease.

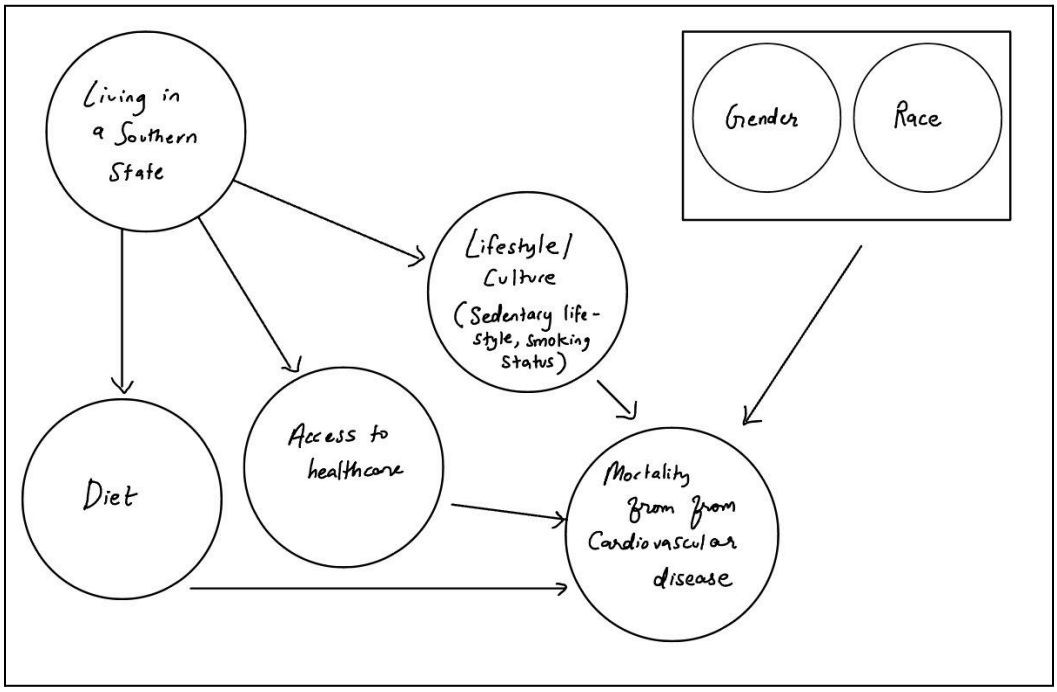


Figure 5: Causal directed acyclic graph

Casual DAG for determining causality between living in a Southern state and mortality from cardiovascular disease. Includes several confounders, only one of which we controlled for - race.
Results

We used three methods of causal inference - Propensity Score Matching (PSM), Inverse Propensity Weighting (IPW), and IPW with Trimming (20%). PSM resulted in an estimated treatment effect of 766. This means that an estimated 766 more individuals died from cardiovascular disease in Southern states. The treatment effect from IPW was 2746.67, suggesting that approximately 2747 more people died in Southern states from CVD. We ended up trimming 20%, resulting in a treatment effect of approximately 1837.59 (approximately 1837 more people in Southern states died from CVD).

We conducted a significance test with the null hypothesis stating that there is no difference in mortality between Southern states and non-Southern states and the alternative stating that there is a difference. The p-value was 0.426 so we failed to reject the null. Therefore, there is no significant difference between mortality rates from CVD between Southern states and non-Southern states.

The uncertainty in our estimates comes from the fact that the confoundedness assumption did not hold, i.e. all confounders were not controlled for. The most obvious one was gender, but other confounding variables affect whether or not a person dies from CVD or not - are they a smoker? Do they live in a food desert? Are they obese? Are they over the age of 50? Do they have a sedentary lifestyle?

As discussed previously, the dataset lacked values for some columns or years. We also did not know exactly how the CDC collected this data - how frequently and from where. All of this adds to our uncertainty.

Discussion

The problem we faced was with how the dataset was structured. While we wanted gender as a confounder, the data had separate categories for race and gender, which grouped all females of all races into one and all genders of all races into one. Therefore, we had to choose between the confounders, so we dropped gender and used race. This doesn't account for gender differences in CVD development, particularly the fact that coronary heart disease, a type of cardiovascular disease, is two to five times more common in men than in women in the younger age groups. Although there is growing research to suggest that there's a more prominent increase of CVD risk in women above 50. (Möller-Leimkühler, 2007)

To address the problem above, it would be great to have data specific to gender and race. Rather than the categories ‘male,’ ‘female,’ ‘White,’ ‘API,’ ‘Black,’ and ‘Hispanic,’ it’d be nice to have ‘femaleWhite’ and ‘femaleAPI’ etc. so we can account for both race and gender as opposed to just race. Additionally, it would be helpful to know if the person smoked, their age, and their weight, as these are all factors that contribute to a higher risk of CVD.

We are fairly confident that there is not a causal relationship between our chosen treatment and outcome, as our statistical test from our limited data produced a p-value of over 0.4, which meant we failed to reject our null hypothesis of there being no difference in mortality from CVD between Southern and non-Southern states. This also corresponds with the trends observed in Figure 3, as there was little to no visual difference. We cannot quantify this confidence.

Conclusion

We utilized the CDC's 2023 Release of the U.S. Chronic Disease Indicators (CDI) and supplemental EPA air quality data to investigate the impact of environmental factors on chronic disease mortality and geographic influences on health outcomes. Our research questions focused on mortality data concerning CVD and COPD. During this project, we encountered significant data gaps and uncertainty, which future studies can improve upon.

For our first analysis, we found that NO2 is the only statistically significant pollutant, with a negative coefficient of -0.0654, indicating a slight negative relationship between NO2 and adult COPD mortalities. We initially expected all pollutants to be statistically significant since scientific evidence suggests a positive relationship between the pollutants and adult COPD mortality rates. We assume that these results are a consequence of over-aggregating our CDI and EPA datasets during the pre-processing stage. The nonlinear relationships between our features and outcome, along with collinearity among our features, may play a role as well. In addition, the Random Forest nonparametric model performs far better than the two GLM models, which produced comparable MAEs and R-squared scores. This aligns with expectations, considering that the Random Forest model is designed to be highly accurate. In summary, our first research question indicates that our features (CO, NO2, O3, and SO2 concentrations) lack predictive power with respect to our outcome (adult COPD mortalities).

We confined our analysis to the top 25 most urban states (per the Census Bureau). This focus inherently limits the generalizability of our findings to similar urban settings and may not accurately represent less densely populated states. Moreover, each state's air quality measurement was recorded at a single location. This further decreases our generalizability, as

that location may or may not accurately represent the air quality for the entire state on that specific day/time. Therefore, we deduce that our findings for our first analysis are quite broad.

We merged the EPA's pollution dataset with the given CDI data (COPD) so that we could easily cross-reference the datasets and gain comprehensive insight across multiple data variables. However, this combination could introduce inconsistencies in scale and granularity, oversimplifying complex relationships among our data, which can lead to faulty conclusions.

Our second research question explored the causal relationship between cardiovascular mortality and residency in a Southern U.S. state by using three analytical methods: propensity score matching, inverse propensity weighting (IPW), and IPW with trimming. We obtained a positive treatment effect (766, 2747, and 1836 respectively) with all three methods. However, we failed to reject the null (p-value = 0.426) when using a t-test to test whether there was a difference in cardiovascular disease mortality numbers between Southern states. Therefore, we fail to conclude that there was a significant increase in deaths from CVD in Southern states.

For this question, our investigation is specific to the geographical and demographic characteristics of these states. Our analysis only pertains to the causal effects within the year 2020 and did not stratify into more specific types of cardiovascular disease or specific demographics like gender or race. The generalization of these results beyond the dataset or similar settings is constrained because of our narrow scope.

In general, the CDI dataset suffers from low granularity, poor documentation, and ambiguity of several variables that pose challenges, complicating the interpretation and the reliability of our analysis. Furthermore, the inability to fully adjust for all relevant confounders in our data analyses emphasizes the challenge of asserting strong causal relationships in observational studies. It is impossible to account for these limitations because we did not

participate in the data collection process or have access to the methodology of the original researchers. Conversely, we did not encounter any limitations for the EPA dataset.

Because of these limitations, the strength of our statistical results and conclusions is not ideal. Regardless, we urge the CDC and state governments to exhaustively collect and maintain important datasets for data science while reinforcing guidelines that safeguard human contexts and ethics. In addition, we recommend that more stringent documentation standards be upheld within federal and state data collections to ensure more accurate and powerful conclusions can be made on these life-threatening chronic diseases.

Although our results for our first analysis may have been unexpected, future studies can build on our work in our first analysis by finding or collecting air quality data in multiple locations across a single state. That way, we can obtain an aggregation more representative of the entire state instead of a single location. Moreover, further examination can be conducted regarding the relationship between the pollutants to reduce sources of collinearity, bolstering the strength of GLM and nonparametric models. Future studies can also build on our second analysis by controlling for the various confounding variables we found when conducting causal inference. A simpler approach would be to control for a smaller cluster of Southern and non-Southern states (i.e. 5 each) that are similar in size and population. Differences in food deserts, lifestyle, access to healthcare, and consequent effects on mortality from cardiovascular disease will be easier to spot when narrowing down our data analysis.

Our main takeaway is that it is difficult to establish causality in an observational study where controlling for all confounders is impossible. Moreover, the convergence of datasets with varying scales and granularity should be carefully monitored as the over-aggregation of data when merging datasets can lead to losing valuable contextual information. The quality,

granularity, and documentation of the dataset have the greatest impact on the generalizability and confidence of derived conclusions. In closing, the opportunity to practically apply learned techniques and concepts to unprocessed, real-world datasets was invaluable.

Dataset Sources

Chronic Disease Indicators Dataset (CDC):

https://chronicdata.cdc.gov/Chronic-Disease-Indicators/U-S-Chronic-Disease-Indicators-CDI-2023-Release/g4ie-h725/about_data

U.S. Pollution Data (EPA):

https://www.kaggle.com/datasets/sogun3/uspollution?resource=download

Works Cited

"America's Most Urban States." America's Most Urban States | Newgeography.Com, www.newgeography.com/content/005187-america-s-most-urban-states. Accessed 6 May 2024.

Centers for Disease Control and Prevention. (2024, March 11). Data Access - National Death Index. Retrieved from https://www.cdc.gov/nchs/ndi/

Centers for Disease Control and Prevention. (2015, January 9). Indicators for Chronic Disease Surveillance - United States, 2013. Retrieved from https://www.cdc.gov/mmwr/preview/mmwrhtml/rr6401a1.htm

Centers for Disease Control and Prevention. Interactive Atlas of Heart Disease and Stroke. Retrieved from https://nccd.cdc.gov/DHDSPAtlas/?state=County&class=1&subclass=1&theme=1&filters=%5B%5B9%2C1%5D%2C%5B2%2C1%5D%2C%5B3%2C1%5D%2C%5B4%2C1%5D%2C%5B7%2C1%5D%5D&ol=%5B10%2C14%5D

Centers for Disease Control and Prevention. (2016, January 4). NVSS - about the National Vital Statistics System. Retrieved from

https://www.cdc.gov/nchs/nvss/about_nvss.htm#:~:text=These%20data%20are%20provided%20through,data%20are%20also%20available%20online

Möller-Leimkühler AM. “Gender differences in cardiovascular disease and comorbid depression.” Dialogues Clin Neurosci. 2007;9(1):71-83. doi: 10.31887/DCNS.2007.9.1/ammoeller. PMID: 17506227; PMCID: PMC3181845.

https://www.simplilearn.com/tutorials/data-science-tutorial/random-forest-in-r

Office of Minority Health. Heart Disease and Asian Americans. Retrieved from https://minorityhealth.hhs.gov/heart-disease-and-asian-americans#:~:text=Overall%2C%20Asian%20American%20adults%20are,to%20die%20from%20heart%20disease