Data Overview

Dataset 1: U.S. Chronic Disease Indicators: Asthma

Dataset 2: Daily Census Tract-Level PM2.5 Concentrations, 2011-2014

Dataset 1 was generated by combining data regarding hospitalizations, mortality rates, patient diagnoses' and hospital visits from 4 sources: Behavioral Risk Factor Surveillance System (BRFSS), State Emergency Department Databases (SEDD), State Inpatient Database (SID) and National Vital Statistics System (NVSS). BRFSS data represents a sample, and was generated using phone surveys. SEDD and SID data are samples generated using hospital files and records. NVSS data is generated through civil registration, an administrative system used by governments to record vital events in their populations. On the whole, dataset 1 is a *sample* because it combines many data sources, some of which consider samples instead of a census. For example, the BRFSS is a phone survey which considers only some individuals, so the overall data is examining only part of the population to represent the whole.

Dataset 2 was generated by combining PM2.5 monitoring data from the US EPA repository of ambient air quality data and simulated PM2.5 data from the deterministic prediction model. It is a *census* because the geographical scale and scope is "all census tracts in the contiguous United States" (*CDC*). For Dataset 1, we found that the distribution of our race variable was limited when we filtered for results corresponding to asthma-mortality rates, as opposed to other asthma-related rates and questions. Our race variable only including White, Black, API, and Hispanic individuals, and excluded American Indian Alaskan Native, Mixed Raced, and Other races. For Dataset 2, non-contiguous US census tracts—Hawaii and Alaska—were excluded, which could exclude Pacific Islanders and Indigenous communities populous in those areas. For the other 48 states, no groups were systematically excluded.

With regards to participant awareness for data sources from Dataset 1: BRFSS is a telephone-related survey, so participants were actively engaged and aware of the collection and use of this data. NVSS is an inter-governmental data sharing of public health information. Although the level of active awareness about NVSS data collection is lower than BRFSS, it is commonly known that patient health information will be shared at an inter-governmental level. SID, SEDD are state-specific files that contain all inpatient care records or emergency hospitalization records in participating states. Their levels of data awareness are similar to NVSS, as inpatient and emergency hospitalization data are part of public records. In conclusion, the level of awareness of data usage and collection depends on the data type. For Dataset 2, participants weren't actively aware of the collection or use of this data because the data was location-specific (not individual-specific) and hence automatically monitored by the EPA.

Regarding granularity, each row in Dataset 1 represents a unique answer to an asthma-related health question in a specific year, for a specific state, based on a single stratification method and reporting unit. This impacts the interpretation of our findings because there are various combinations of question/answer sets we can explore, adding more specificity. Thus, we can get a sense of asthma mortality rates for various combinations of year, race, gender, location, and data type values. For Dataset 2, each row represents a different US census tract in 2011-2014, and its corresponding PM2.5 levels, among other variables. This impacts our interpretation of findings because the boundaries of a census tract may not represent demographic variations within and across different tracts, and because we can only make inferences about PM2.5 levels at the tract or larger levels of location aggregation.

There are some concerns relevant in the context of our data. Selection bias occurs in Dataset 1 as SEDD and SID exclude populations such as veterans, Native Americans, and institutionalized populations, while BRFSS exclude college campuses or individuals in the military. As a result, our study population is not representative of the target population. In addition, measurement error may be present since asthma mortality rate is measured through death certificate data but the reliability of such data has

been questioned, particularly for older groups. Lastly, BRFSS is an example of convenience sampling as it excludes individuals who do not own a cell phone. For Dataset 2, selection bias or convenience sampling is not present, but measurement error may arise as AQS is not frequently updated (Sanchez).

There are important features we wished we had to answer further questions. For Dataset 1, stratification based on people with underlying health conditions, such as cardiovascular disease or diabetes, would be helpful. This would help us understand asthma mortality rates in conjunction with underlying diseases. For Dataset 2, information regarding the size of each census tract would be helpful. This would help us understand the concentration levels of pollutants in relation to how many people are impacted by them. Pollution concentration levels for other pollutants such as O3, PM10 and PM0.1 may also help us get a better sense of the causal relationship between various pollutants and asthma mortality rates, as opposed to the specific causal relationship between pollutant PM2.5 and asthma mortality rates.

Research Questions

Our first research question uses causal inference to explore: **Does living in states with higher air pollution levels cause an increase in asthma mortality rates?** By answering this question, we can understand how pollution levels affect asthma mortality rates, as well as determine which geographical locations in the US may be the most and least safe for individuals with asthma. Causal inference is a good fit for this question because the question draws a direct comparison between air pollution levels and asthma mortality rates. Thus, we can understand how changes in air pollution, as mediated by state location, cause changes in asthma mortality rates and draw causal rather than associative relationships between our variables. To test the strength of such causal quantities, we observe the associated p-values.

Our second research question compares GLMs and nonparametric methods to explore: **How well does race predict risk for asthma mortality?** By identifying which races are the most at risk for asthma mortality through GLMs, we can help decide which racial and ethnic populations need more

healthcare, resources, and structural support for asthma care. Thus, our findings guide decisions for policy and public health intervention to combat racial disparities in asthma mortality prevalence, as well as the underlying social determinants of these disparities. Utilizing GLMs is a good fit for our question because we can model our continuous response variable as asthma mortality, and our predictor variable as race via Gaussian linear regression. GLMs are flexible generalizations of linear regression, predicting an unknown response variable as a linear combination of observed predictors (Ram).

EDA

Asthma Mortality Rates by State: Figure 1

We observe that states in the North (South Dakota, Illinois, New York) tend to have higher asthma mortality rates than states in the South (Texas, Kentucky, Florida). Although this trend is not always consistent and there are outliers (e.g. Mississippi in the South having a high mortality rate), this trend can be observed for a majority of the states. After this, we would want to identify potential differences in asthma mortality rates in rural, urban, or suburban areas. We would also want to look at other demographic or geographic factors to see how they cause differences in asthma mortality by state.

In terms of data cleaning, we filtered out rows that had a null value for the Asthma Mortality prevalence column. We also removed rows that had a location description of “United States,” to ensure that location is considered on a state-wide basis and granularity is not lost in our model. Furthermore, we only looked at values which had the age-adjusted rate – rather than the raw number or the crude rate . This impacts our model because it removes the confounding effect of age from our model.

Through this visualization, we’re able to effectively identify states with higher asthma mortality rates than others. This motivates questions about why some states have higher rates of asthma mortality than others – is it due to those states having higher air pollution levels?

Asthma Mortality Rates by Race: Figure 2

One trend we observed was that the Black community had a way higher average age-adjusted mortality rate than any of the other three races we looked at (API, Hispanic or White). We may want to follow up on the relationship of asthma mortality rates stratified by gender or other demographic factors WITHIN race; i.e. are women of a certain race more likely to die from asthma than men of that race?

Regarding data cleaning, we filtered out null values in the data value column, and only looked at values relevant to asthma mortality rates. We also looked at only those values which had the age-adjusted rate, rather than the number or the crude rate, which removes the confounding variable of age from our model. Because we only focused on the question of asthma mortality, we were able to gather information to build our model only on mortality rather than prevalence and hospitalization rates.

While we were exploring demographic factors such as state location, we realized that it would be helpful to consider other factors such as race and their relationship to asthma mortality. This allows us to analyze how well race predicts risk for asthma mortality, maybe even in conjunction with gender. Our data suggests that racial minorities tend to have a higher risk for asthma mortality than racial majorities. Perhaps this is due to differences in socioeconomic status and lack of access to better healthcare.

Asthma Hospitalization Rates by Race: Figure 3

We noticed that the Black community has a much higher rate of asthma related hospitalization than all other racial groups, followed by American Indian Alaskan Native individuals, Hispanic individuals, White individuals, and then API individuals. This trend seems to follow that of asthma mortality rates by race - minorities have higher rates of asthma mortality; however, it is interesting to note that Asian-Americans, rather than Whites, have the lowest rates of asthma hospitalization. We would want to follow up on the stratifications of race and socioeconomic status for asthma hospitalization rates.

To clean our data, we filtered out null values for the data column, and only looked at values relevant to asthma hospitalization rates. We also looked at only those values which had the age-adjusted rate, rather than the number or the crude rate, which removes the confounding variable of age. Because we only focused on the question of asthma hospitalization, we were able to gather information to build our model solely on hospitalization rates rather than prevalence and mortality rates.

Our visualization is relevant to our research question of how race predicts risk of asthma mortality, as it provides insights into how race predicts risk for asthma hospitalization, an adjacent event that could serve as an indicator for asthma mortality rates. Thus, our visualization motivates a larger question of how asthma related health events, such as mortality rates, differ and are predicted by race.

Asthma Mortality Rates versus Pollution Levels: Figure 4

We notice that there is a weak positive correlation between pollution rates and age-adjusted asthma mortality rates. Although this result is positive as one would expect (i.e. higher pollution rates lead to higher age-adjusted asthma mortality rates), the correlation is not as strong as we would expect it to be. This motivates us to follow up on whether confounding variables that may influence both pollution and asthma mortality rates exist. In terms of data cleaning, we only looked at rows that had a non-null state ID. Not including these in our model prevents the introduction of outliers that would make incorrect inferences. We also removed values with a "United States" location description to ensure that location is considered on a state-wide basis and granularity is not lost in our model. Finally, we looked at only those values for age-adjusted rate, removing the confounding variable of age from our model.

This visualization motivates our question because it introduces the relationship between pollution rates and age-adjusted asthma mortality rates. However, it does not suggest a potential answer because the two variables don’t have a strong correlation, preventing us from drawing reasonable

conclusions. Nonetheless this prompts us to think about how confounding variables like access to healthcare and regional median income affect asthma mortality rates.

Modeling Techniques

Option C - Prediction with GLMs and nonparametric methods

Methods

We're trying to predict an individual's risk of asthma mortality, or asthma mortality rate, based on race. Using *pd.get_dummies*, we were able to one-hot encode the race variables of White, Hispanic, Black, and Asian and Pacific Islander and translate that into the features of our model. We decided to use a Frequentist GLM with a Gaussian likelihood to conduct linear regression with asthma mortality as the response and race as the predictor variable, because we believe the two have a linear relationship. Our GLM operates under the assumption that asthma mortality rate is a continuous, unbounded variable. We're also assuming the racial categories for our predictor variable are mutually exclusive and collectively exhaustive. For our nonparametric method, we chose Random Forests, because they can handle complex relationships among features. Additionally, they perform better on test sets than decision trees by combining decision trees to limit overfitting. We make the assumption that our one-hot encoded race variables generate a sufficient number of features necessary for node splitting. Here, we're assuming racial categories for our predictor are mutually exclusive and collectively exhaustive.

We evaluated performance of our random forest via RMSE and accuracy, and GLM performance via goodness-of fit-checks: how close the log-likelihood is to 0, as well as how close deviance and chi-squared values are from n-p: n is the number of observations, p is number of parameters.

Results

For our non-parametric method, we fit our random forest and ended up getting a 67.04% accuracy for our training set and a 75% accuracy for our test set. This suggests that race is an important feature to predict risk for asthma mortality, but there probably exist other features that may be working

in conjunction with race to predict asthma mortality. We should consider these features in order to achieve higher accuracy. For our GLM, we observed that the race with the largest positive coefficient was Black, and the race with the largest negative coefficient was White. This suggests that being Black is likely associated with a significant increase in risk of asthma mortality, whereas being White is very unlikely to result in a significant increase in risk of asthma mortality. There is also a large intercept term which is suggestive of other "noise" affecting risk of asthma mortality rates.

The coefficients for Black and White are statistically significant because their p-values (0.000 and 0.001 respectively) are less than the significance threshold of 0.05, whereas the coefficients for Asian Pacific Islander and Hispanic are not (0.442 and 0.614 respectively). This suggests that there may be room for doubt when predicting asthma mortality risk for individuals who are of an Asian Pacific Islander or Hispanic identity, but there is stronger evidence for a relationship between an individual's White or Black identity with their risk for asthma mortality rate. To estimate uncertainty in our GLM, we used bootstrapping to generate 95% frequentist confidence intervals for each regression coefficient. Bootstrapping works well because the number of parameters (5) is less than the number of data points the parameter depends on (326). We found that the 95% bootstrapped confidence interval was (10.35, 11.7) for the constant, (-2.49, -0.28) for API individuals, (12.45, 15.12) for Black individuals, (-1.77, 2.73) for Hispanic individuals, and (-2.56, -1.17) for White individuals. Based on the confidence interval size, we are most certain about our estimates of asthma mortality for White and Black individuals, and least certain about our estimates for API and Hispanic individuals, which aligns with our findings.

Discussion

GLM performed better because the bootstrapped standard errors for coefficients are very small, and the coefficients are statistically significant. Even the largest error (1.16 for the Hispanic coefficient), is substantially lower than the decision tree RMSE. Furthermore, it's a more interpretable model to show how race predicts risk of asthma mortality, by showing how strongly each racial group predicts risk via

regression coefficients. The only caveat in terms of goodness-of-fit for the GLM is that the deviance (8345) and chi-square (8.35e+03) are not close to the difference between the number of data points (326) and the number of parameters (5), or 321. That being said, we are confident in applying this to future datasets because we took census data and stratified across races; the census can be generalized to represent findings for the US population.

For the random forest, we found that White individuals had a predicted asthma mortality rate of 9.13%, Black individuals had a predicted rate of 24.95%, Hispanic individuals had a rate of 11.46%, and API individuals had a rate of 9.65%. The predicted mortality rate is highest for Black individuals, followed by Hispanic, API, and and White. This is consistent with our expectation. From our GLM model, we found that the coefficient for increase in mortality rate for Black individuals was 13.81, API is -1.41, White is -1.9, and Hispanic is 0.5. This means that if all other variables were held constant and an individual was Black, for example, their risk for asthma mortality would be increased by 13.8%.

The limitations of our models include potential bias in collecting information on different races. Since it has been shown that minority races don't respond to census as much (*The Guardian*), there may be bias in our model. Furthermore, our model does not take into account how people of multiple races are reporting themselves, which may lead to inaccuracies in our model due to the ambiguity of data. This is because when we filtered for data about asthma mortality rates with non-null data values, we lost information on individuals identifying as Mixed Race, Other, or American Indian Alaskan Native.

Some additional useful data for improving our model would be stratifying by income or socioeconomic status alongside race; people in a lower economic class might be at higher risk for asthma mortality, particularly for low income individuals of color. We could also look at data on family history of asthma, or other health conditions that could affect individual risk of asthma mortality.

Option D - Causal Inference

Methods

For our causal inference question, the treatment variable is pollution rates (i.e. the concentration of PM2.5), and the outcome is asthma mortality rates. A confounding variable is state location, which has a causal relationship to both pollution and asthma mortality rates. The degree of industrialization and environmental consciousness on a state-wide level directly impacts pollution levels. State location also affects asthma mortality rates as quality of healthcare and age demographics differ by state. To adjust for confounders, we controlled for state location in our model. Our data shows that California, New Jersey, Texas, New York and Florida have the most available data points, so we only analyzed the causal effect of pollution rates on asthma mortality rates in those states. This allows us to only look at states which have a similar degree of industrialization and environmental consciousness, relative to other 50 states, because of their large population size. Additionally, there are no colliders present in the dataset.

Results

The causal effect of pollution rate on asthma mortality rates in all states is ~0.002 without adjusting for confounders. Hence, if you increase pollution rates by 1 unit and hold everything else constant, asthma mortality rates increase by 0.002 units. Thus, we can infer a weak but positive causal relationship between pollution and asthma mortality in all states. The causal effect of pollution rates on asthma mortality rates in the most populous, industrialized states (California, New Jersey, Texas, New York and Florida) is around -0.26 after adjusting for confounders. This means that if you increase pollution rates by 1 unit and hold everything else constant, asthma mortality rates actually decrease by ~0.26 units. We did not assume unconfoundedness in our model, which led us to adjust for our confounders. The assumption we made was that California, New Jersey, Texas, New York and Florida were indeed the most populous states, which we assumed meant they were the most industrialized and highly polluted. However, our data shows a negative causal relationship between pollution and asthma

mortality rates in these states; perhaps this is because these states have more widespread healthcare capabilities and resources to combat exposure to pollution, because they are larger and more urbanized.

We know that if the likelihood of observing our data given the coefficients is close to 1, the log likelihood should be close to 0. The log likelihood of our model when considering all states is fairly large (-95.36), but gets much closer to 0 when we select for the most populous and industrialized states (-12.21). Thus, the goodness of fit of our model, an indicator of model performance, increases after adjusting for the confounder – state location. Regarding uncertainty, we quantified uncertainty by finding standard errors of the parameter estimates from our OLS model. The standard error for the PM 2.5 variable increased, from 0.21 to 1.07, after selecting only the most populous states, meaning that our model certainty decreased. Thus, we know that our method of selection for states should be more robust and generalizable to comprehensively examine the relationship between pollution and asthma mortality.

Discussion

The limitation of our method is that we may not be adjusting for confounders in the most accurate way possible; we assumed that the most populous states would be the most industrialized and thus have the highest rates of pollution, yet did not have as much robust evidence to support this claim. If we had evidence supported by data to know what states were most industrialized and populous, we could adjust for our confounders in a more data-driven way, i.e. national environmental rankings of states. This would be preferable to our current method of adjusting for confounders, which relies on background knowledge about states and their levels of environmental consciousness and pollution.

In terms of statistical inferences, we are not confident there exists a causal relationship between pollution and asthma mortality rates. This is because our p-values are not statistically significant and our log-likelihood is far from 0 for all states and the most populous states, highlighting low goodness-of-fit.

Conclusions

From our causal inference problem, we found no statistically significant causal relationship between concentration of PM2.5 and asthma mortality rates for all states, as well as for the most industrialized ones. From comparing GLMs to non-parametric methods, we found that being Black is associated with a statistically-significantly increased risk for asthma mortality, whereas being White is associated with statistically-significant lower risk. Since we used datasets from the census, our results are very generalizable, making our conclusions applicable to the continental U.S. Our findings made minimal assumptions surrounding population and environmental-consciousness of particular states. However, if we conclude these assumptions are correct, our findings are broadly correct. Notably, our data is from 2011-2014, meaning it may be less generalizable to current phenomenon and contexts.

In terms of calls to action, since we found a statistically-significant, high disparity in asthma mortality risk for Black and White populations, our findings motivate the expansion of resources and research on the adverse effects of pollution amongst minorities. Furthermore, our results suggest a call to action in terms of investigating the social determinants of racial disparities in asthma mortality, such as environmental racism and poor neighborhood environmental conditions in racially underserved areas.

To generate our data for question 1, we merged data gathered from the Chronic Disease Indicators dataset as well as data on PM2.5 concentration levels dataset together. This allowed us to track different metrics for the same state, and see how various quantitative variables like PM2.5 concentration levels affect demographic factors like race, which can predict asthma mortality.

One limitation with our data is that dataset 2 didn't account for information from Hawaii or Alaska; thus, we can't generalize our findings to the entire US population. Additionally, dataset 1 may include measurement errors and convenience sampling. Some future studies that could build on our work would be diving deeper into other types of pollutants that affect asthma mortality rates, aside from PM2.5. This would be helpful when targeting solutions to reduce specific pollutants in the environment.

Further studies could also detect confounding variables that may affect asthma mortality rates, and other demographic factors in conjunction with race that could predict asthma mortality – i.e. gender, income.

Figures

Figure 1

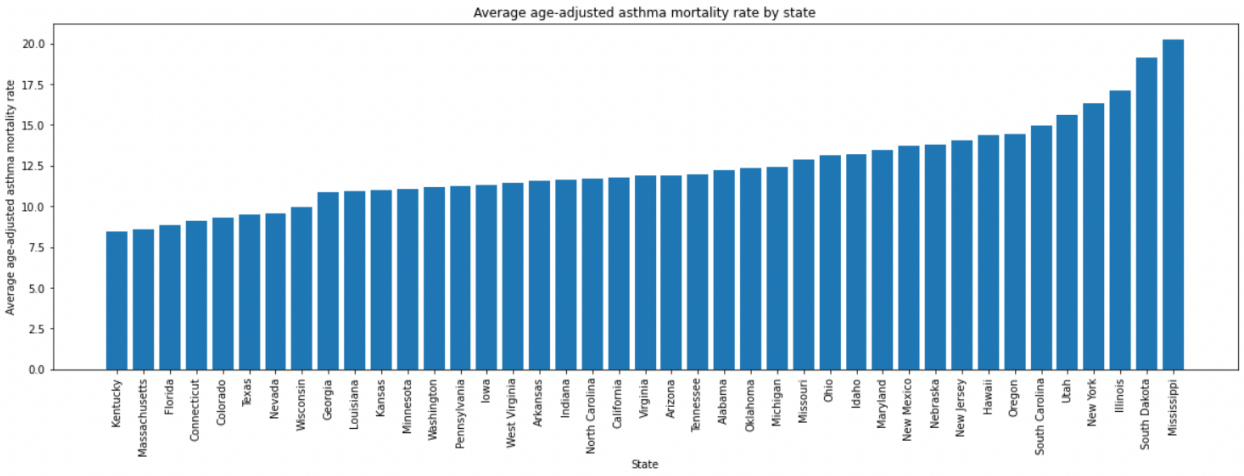


Figure 2:

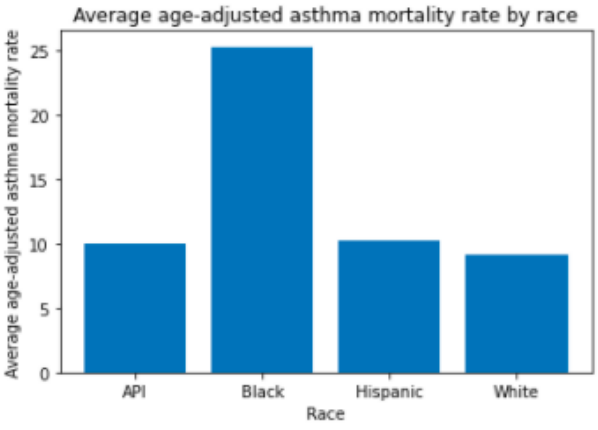


Figure 3:

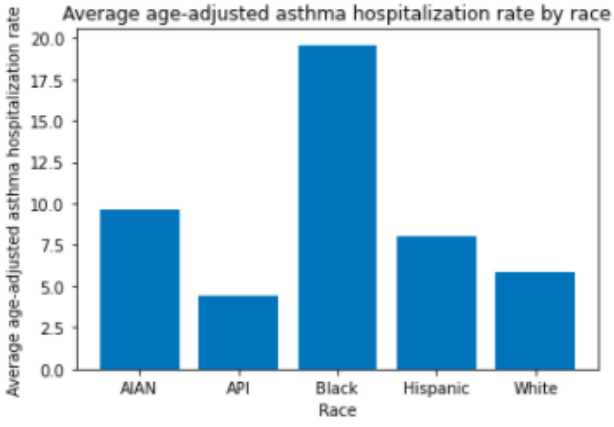
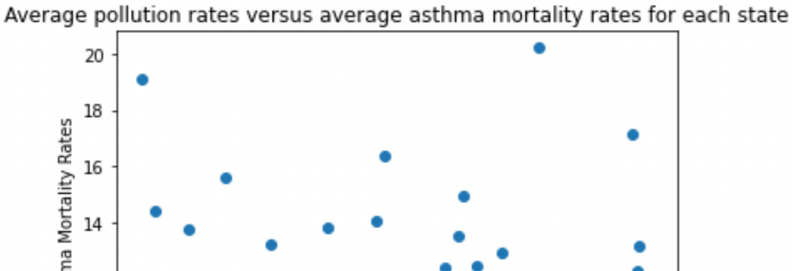


Figure 4:



Works Cited

"How the US Census Misses People of Color – and Why It's so Harmful." *The Guardian*, Guardian News and Media, 27 Feb. 2020, https://www.theguardian.com/us-news/datablog/2020/feb/27/2020-us-census-black-people-mistakes-count.

"Indicators for Chronic Disease Surveillance—United States, 2013 ." *Morbidity and Mortality Weekly Report*, Centers for Disease Control and Prevention, 9 Jan. 2015, https://www.cdc.gov/mmwr/pdf/rr/rr6401.pdf.

Ram, Prabhu. "Generalized Linear Models: What Does It Mean?" *GreatLearning*, 27 Apr. 2021, https://www.mygreatlearning.com/blog/generalized-linear-models/.

Sanchez, Kait. "Go Read This Report about EPA Failures in Air Quality Monitoring." *The Verge*, The Verge, 2 Dec. 2020, https://www.theverge.com/2020/12/2/21833693/air-quality-index-monitoring-epa-pollution-health-reuters.