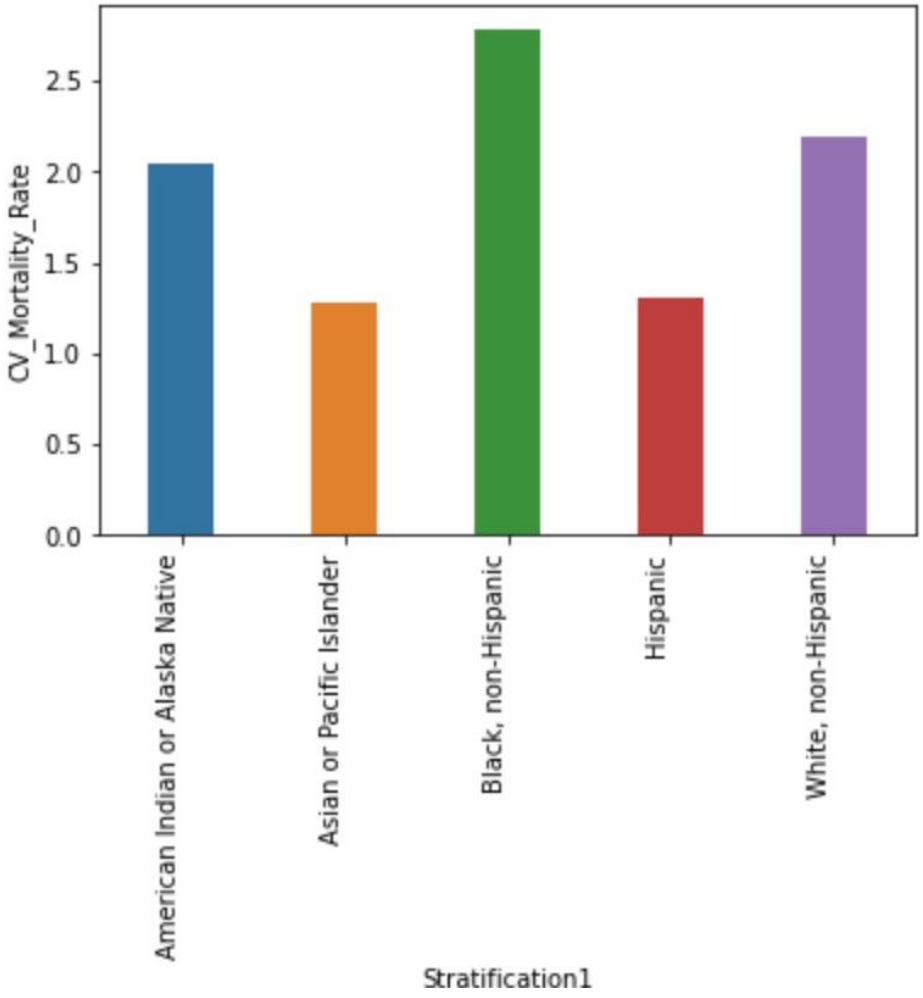
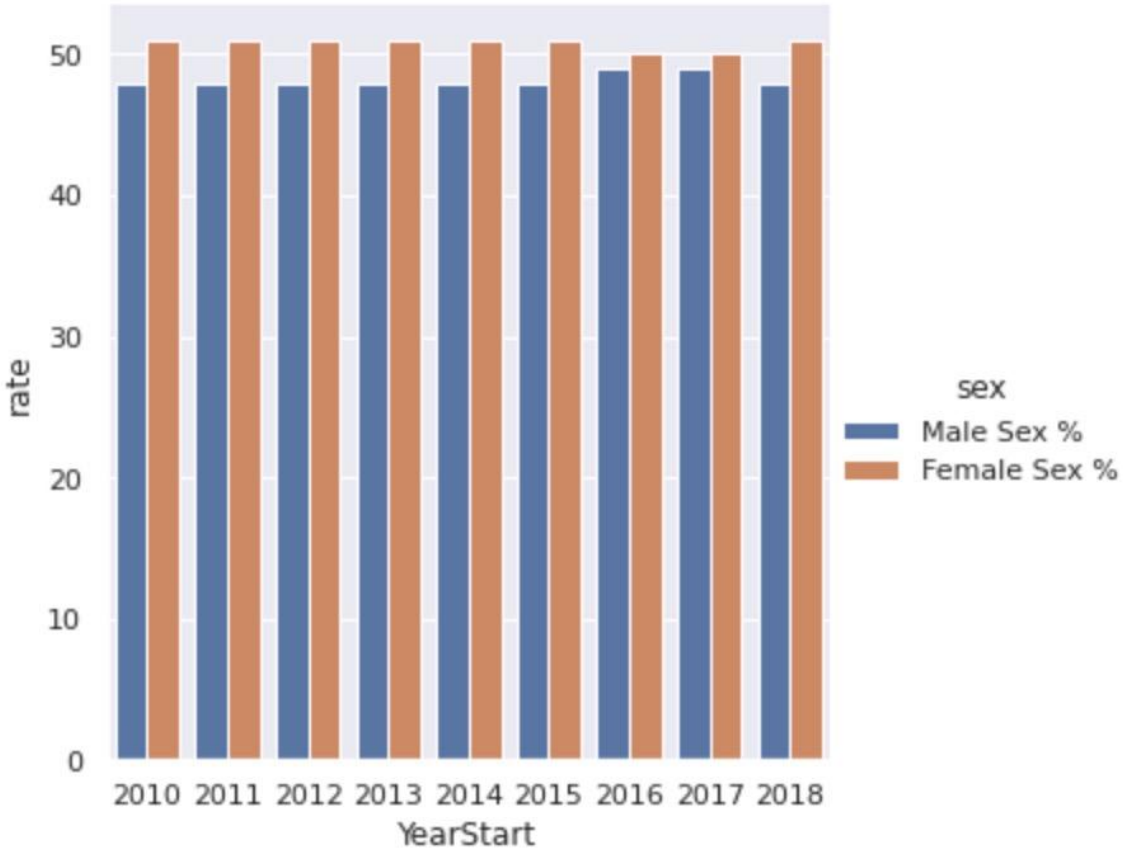
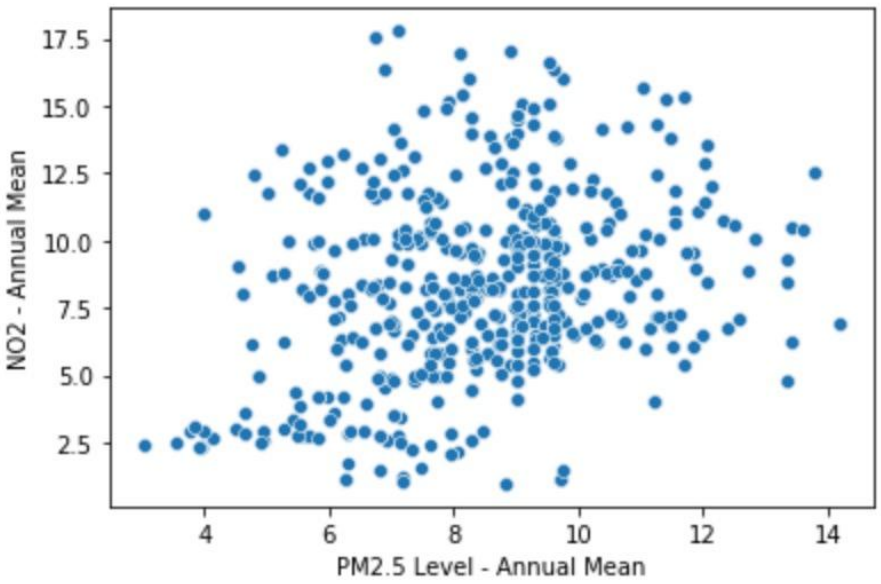
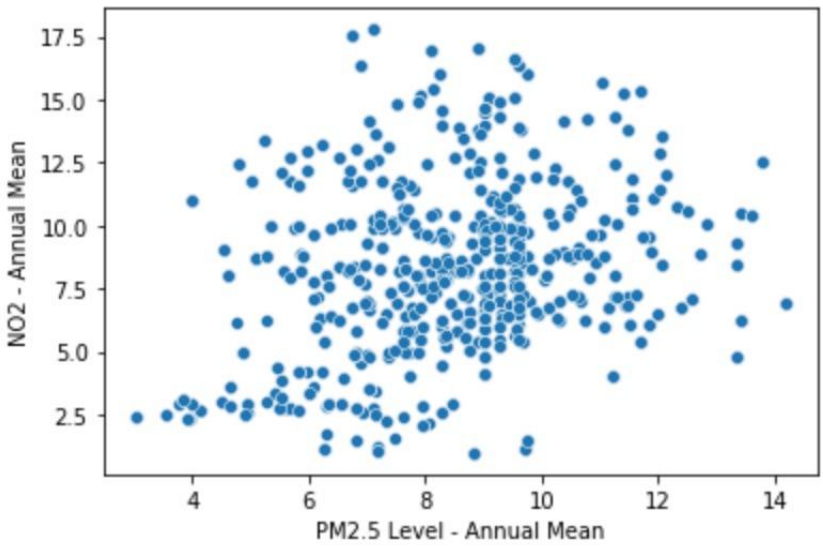
Team Members: Ava Taghizadeh, Jason Chen, Monse Lopez, Sandra Zavala

EDA:

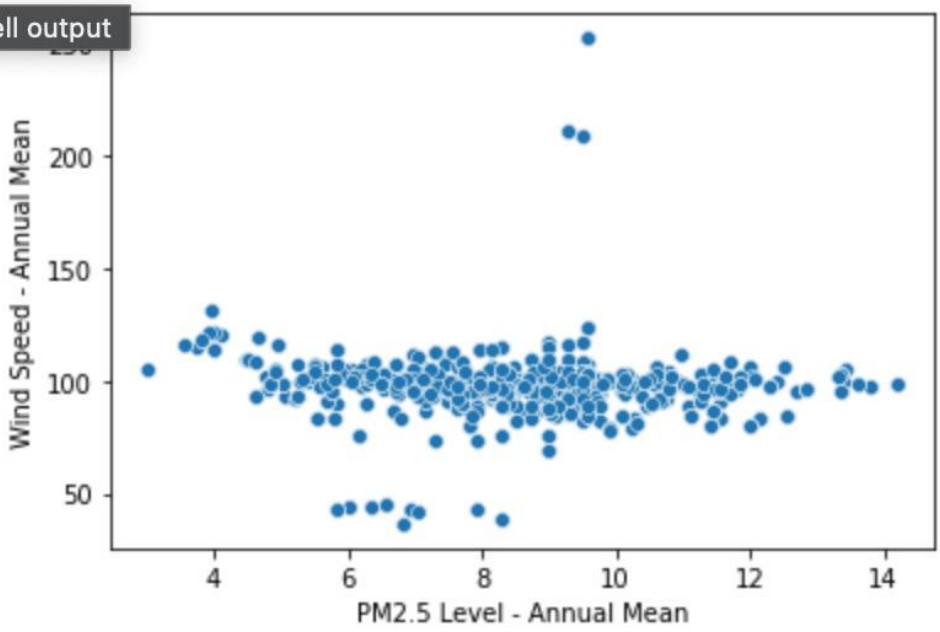



50
40
30
20
10
0
rate
2010
2011
2012
2013
2014
2015
2016
2017
2018
YearStart
sex
Male Sex %
Female Sex %


17.5
15.0
12.5
10.0
7.5
5.0
2.5
NO2 - Annual Mean
4
6
8
10
12
14
PM2.5 Level - Annual Mean



Cell output



Can we predict survival outcomes from cardiovascular diseases (as measured by mortality rates) given certain demographics of an individual? (Comparing GLMs and non-parametric methods):

Methods

Our model seeks to predict crude mortality rates from cardiovascular diseases in the United States using demographic factors such as race and location of residence. Based on the CDC's Annual State-Level US Chronic Disease Indicator dataset, we were able to obtain

age-adjusted mortality rates (cases per 100,000) for this patient population across all states between 2008 and 2020. We picked age-adjusted mortality rates as the data did not include further stratification by age, making it difficult to use age as a feature.

The data provided annual rates on aggregate and further broken down by race/ethnicity for each given year. Given the categorical nature of race, we performed one-hot encoding on our race variable to transform it into usable features. We selected race as a feature because there has been widespread evidence regarding differences in mortality rates among different racial groups due to varying levels of access to healthcare, discrimination, average income levels, etc. For example, it is observed that SES better predicts diseases within the white population compared to other racial groups (Eileen M. Crimmins, Mark D. Hayward, and Teresa E. Seeman). Using race in our analysis may allow us to identify groups that may receive unequal prevention, screening, or treatment within the US healthcare system *(*Scarlett S. Lin and Jennifer L. Kelsey).

Upon further research, we found a research study conducted in Finland that demonstrates how the risks of cardiovascular diseases differ between men and women, which notably changes by age (Mikkola). Thus, we added gender as a feature in our dataset. To account for the improvement of technology and healthcare over time, we also included the year of the statistic to take into account any temporal differences that may be occurring. Based on exploratory data visualizations, we saw that age-adjusted mortality rates were declining over time, which further supported our intuition of including the year as a feature.

To account for state-level variations without one-hot encoding states to have 52 columns, we used several proxies to account for state-level differences in mortality rates. We used per capita medicare expenditure, rate of diabetes, rate of obesity, and rate of tobacco consumption. We utilized per capita medicare expenditure due to the fact that states with higher per-capita medicare expenditure may be able to improve their residences' overall health which improves survival outcomes for all diseases. Additionally, state-funded programs may provide faster and more widespread access to healthcare for lower-income individuals which allows patients to be diagnosed at an earlier stage, further improving outcomes. Rates of diabetes, obesity, and tobacco consumption were included since they are known risk factors for cardiovascular disease to improve our model's accuracy. To handle missing years in some of the datasets, we calculated the annual growth rate for each state and used that to project rates for the missing years.

Since mortality rate is count data, we wanted to use GLM models with poisson or negative binomial distributions. The mean and variance of our mortality rates were not close to each other; therefore, a poisson distribution model was ruled out. We took a frequentist approach here with our GLM models and tested out if our assumption regarding the negative binomial distribution nature of the data held up. We used chi-squares, log-likelihood, and standard errors to evaluate our model choice with a variety of features.

We evaluated our data using decision trees and random forests with different combinations of our features to better understand the impact of each. We used nonparametric models to avoid making assumptions about the underlying distribution of the data.

In evaluating all our models, we decided to use RMSE as our performance metric, since the dependent variable (age-adjusted mortality rates) is continuous data. Given the data reported a lower and higher confidence interval for the mortality rates as their own columns, we compared our RMSE metric to those variations in confidence intervals. We also split our data into a training and a test set to evaluate our models' performances in both the train and test sets.

Results

Nonparametric:

Decision Trees: The first decision tree model only used the one-hot encoded columns of the race feature to predict mortality rates. The train set error was 50.21 and the test set error was 46.25. Given the average range for the lower and higher confidence limit was 45.60, we believed the errors were reasonable. Modeling the decision tree with various features (per capita medicare expenditure, race, gender, year, diabetes, and obesity rate) resulted in a train set error of 0 and a test set error of 48.15, which can be an indicator of overfitting the dataset. As a result, we concluded that the decision tree with only using race was our best model.

Random Forests: Given the positive results of the decision tree, we additionally created a random forest model to evaluate our features. The random forest model using just race as a feature gave us a training set error of 50.21 and a test set error of 45.59. After testing combinations of regressors, we found that the features that gave us the lowest train and test error occurred when using the features: race, medicare expenditure, year, diabetes, and obesity rate. This produced a training set error of 15.29 and a test set error of 36.87.

Since we used decision trees and random forests, interpretation of the models was more difficult, as compared to linear and logistic regression models. However, by performing models on each feature separately, we were able to distinguish between the features that had a larger impact.

GLMs:

We first ran a GLM poisson model with different combinations of features which resulted in overly small standard errors, despite significant p-values. The very small standard errors along with overly narrow confidence rates indicated that a poisson distribution was not a great choice for our data. This further confirmed our earlier conclusion based on the mean and variance of the mortality rates, that a poisson distribution model was not a good fit.

Performing a GLM regression with negative binomial distribution was more successful when using only race as a feature. The standard errors had improved and the model was not as overly confident as before. Additionally, the log-likelihood divided by the number of observations was around -6.25 which was reasonable given that values close to 0 indicate a good model fit. The chi-squared was additionally less than $n - p = 1880$ (given 1885 observations and 5 columns) which supported our other findings. When including other features for this model, our confidence intervals became too wide and the range included negative and positive values, which made the model inference not useful.

In estimating the uncertainty of our GLM predictions, we first looked at confidence intervals. Here, we noticed the issue of overconfidence in our poisson models and the issue of wide confidence intervals for negative binomial distribution models using features other than race (described above in more detail along with log-likelihood and chi-squared). We also looked at p-values (lack of significance when using features other than race) to determine the significance level of the estimations as well.

Discussion

After comparing the results of the models we used, we can see that using Random Forests with the features addressed above produces the smallest error in predictions. We feel confident applying our random forest to future datasets as it was trained using an extensive dataset across 12 years and our errors were within a reasonable range. They can help shed some light on important risk factors and give a general idea of survival outcomes given certain characteristics. However, we would caution against using the models for individual-level diagnosis and

interpretation as they are aggregate-level data and do not take into consideration much important personal-level information, such as age, family history, and genetic disposition.

Both the Decision Trees and the Random Forests were able to provide us with quantifiable evidence of how well these models were performing given our dataset, as well as give us insight as to which features (or combination of features) seemed to have a greater impact on how well our models were predicting. Given the nature of these models and the complexity of them even after using visualizations, it is hard to interpret them to conclude a specific increase in mortality rate given a certain risk factor.

For GLMs, we ultimately concluded that our estimations were not helpful and a linear model was not a great fit for our data. While a negative binomial distribution when using race as a feature seemed to give significant results, most of the features were not utilized and from domain knowledge, we know many of these features are quite significant in mortality risks. Thus, we conclude that a linear model is perhaps not a correct assumption to make for this dataset.

Given our dataset, we are not able to determine the education levels, income levels or age of patients (except for features we included such as gender) that allow us to have a better understanding of the population's socioeconomic status. There is also an issue with using race as a regressor, given that there are cases in which a patient could apply to multiple different races, likewise, it is difficult to capture issues within our healthcare system that may not be reflected in our dataset. For example, black people are less likely to report conditions, such as heart disease, cancer, and chronic lung disease, than white people (*Eileen M. Crimmins, Mark D. Hayward, and Teresa E. Seeman*). We can also not capture the racial disparities in health care. More specifically, in 2005, the National Academy of Medicine released a report documenting that given the disproportionate level of poverty that affects black people causes higher rates of disease and shorter life expectancies than white people. (Khiara M. Bridges). It is important to note that the decision tree model is not ideal as slight changes to input features can have a big impact on the predicted outcome, which is not usually desirable. Trees can also be unstable since a few changes to the training dataset can create a completely different tree. For GLMs, they rely on an assumption of linearity between the response and predictor variables which is a strong assumption to make in our dataset that proved to not hold.

As mentioned above, having more information as to the socioeconomic status of each patient would help us create a model that better fits the population we are using in our regression, thus helping us develop a better prediction for future populations. The progression of a patient's disease given the year of diagnosis may also serve as a valuable feature that could show the disparities minority groups face in the health care system. Similarly, the length of treatment before death could allow us to see if there are any differences in survival rates between races.

Does air quality have a causal effect on crude age-adjusted mortality rates of cardiovascular disease? (Causal inference):

Methods

Our model seeks to interpret whether air quality has an effect on the crude age-adjusted mortality rate from cardiovascular disease. Thus, the treatment is the annual average air quality of that state (as measured by PM2.5 levels) and the outcome is the crude age-adjusted mortality rate of cardiovascular disease. We have used a daily air quality dataset collected by the National

Environmental Public Health Tracking Network that details the amount of PM2.5 levels by census tract along with our dataset from the previous question that includes mortality rates.

There are many confounders when thinking about the effect of air quality on mortality rates from cardiovascular disease. Similar to our reasoning for our previous research question, we are including race as a confounding variable to account for disparities in medical care within particular races that may have an effect on our outcome. A similar logic can also be applied to other factors such as tobacco consumption. If an individual is smoking cigarettes, they are likely increasing their exposure to more PM2.5 particles, while simultaneously increasing their risk of heart disease. As a result, the unconfoundedness assumption does not hold, which makes causal inference for PM2.5 levels on mortality rates difficult. For this reason, we decided to implement the annual mean of NO2 per state as an instrumental variable for our model.

Given that this was not a randomized experiment and was a natural experiment based on real-world data, we used an instrumental variable to help introduce a partial element of randomness to help us with the causal inference. Initially, the wind speed was chosen as an instrumental value in order to account for concentrations of PM2.5 in certain areas after finding a research paper with a similar study that accounted for this feature given that it affects concentrations of local emissions and confirmed that the instrument was not associated with mortality independent of pollution (Joel Schwartz, Marie-Abele Bind, and Petros Koutrakis). Ultimately, we found that wind speeds produced a very large prediction for the volatility PM2.5 levels (the 2SLS estimate of PM2.5 levels being 3482.807) have on mortality rates. We then looked to NO2 levels as an IV we could use for PM2.5 levels. Looking at the correlation between PM2.5 and NO2 levels as well as NO2 and mortality rates gave us the confidence to use NO2 as an instrument variable (which was also used as an instrument in the study we found) and found much more promising results in the OLS model for mortality rate, which produced a 2SLS estimate of PM2.5 levels of …

Given that a collider is a variable that is caused by both the outcome (mortality from cardiovascular disease) and treatment (exposure to poor air quality), no colliders come to mind regarding this dataset since the observed outcome is a mortality rate.

Results

[Summarize and interpret your results, providing a clear statement about causality (or a lack thereof) including any assumptions necessary.]

Ultimately, we found that wind speeds produced a very large prediction for the volatility PM2.5 levels (the 2SLS estimate of PM2.5 levels being 3482.807) have on mortality rates.

[Where possible, discuss the uncertainty in your estimate and/or the evidence against the hypotheses you are investigating.]

Upon further research on the subject of air pollution, we discovered that there may be a multitude of variables that can affect PM2.5 levels in a local area. These factors include (but are not limited to): seasonal temperature, wind speeds, black carbon levels, NO2 levels, the height of planetary boundary layer (PBL; topography). Although there have been studies to suggest a correlation between air pollution and mortality rate, the model we created is not strong enough to capture the complexities of this topic. We developed our estimate with a certain level of uncertainty, given our lack of knowledge in the environmental factors that affect levels of pollution.

Discussion

While using an instrumental variable is a great way to help with causal inference in the case that confounders are hard to account for, finding a good instrumental variable is extremely challenging and a task that requires significant domain knowledge. We used existing research papers to help narrow down our instrumental variables but ultimately had a difficult time as many did not show the correlation relationship that was needed (a correlation between IV and treatment while no correlation between IV and outcome).

Another form of data that can be considered is an exposure measurement error, which is not usually available when in the conversations of defining the relationship between air quality and a disease outcome. This error should be considered as it can impact the inferences resulting from the research question and can help determine how biased the impact of air quality has on a disease outcome (Sheppard).

[How confident are you that there's a causal relationship between your chosen treatment and outcome? Why?]

Based on existing literature, we know that there is a causal relationship between PM2.5 levels and cardiovascular disease. One study from the EHP concluded that there seemed to be an overall increase in mortality with the increase of PM2.5; however, most current literature focuses on the prevalence of the disease as opposed to mortality rates. We were limited by our dataset as it only included mortality rates but believe a causal relationship still exists.

Conclusions

We saw the random forests worked the best for predicting survival rates for patients given certain demographic features, and after regressing our data with each feature, we were able to find that year, gender, race, medicare rates, diabetes rates, obesity rates within each state in the US gave us positive results in fitting our mode. Using a combination of all these features allowed us to create a model that predicts mortality rates, with a small level of error in the training and test sets-- given certain confidence intervals. Regardless of this model, we did find in researching this issue that there are many biases we are not able to completely account for and unobservable disparities that continue to occur within the healthcare system.

We had a similar conclusion within the causal inference we performed on mortality rates given the level of pollution in a particular area. Although we were able to find an instrumental variable (NO2) that was correlated with PM2.5 levels, this correlation was still weak. We found that developing a model that we could confidently use in predicting mortality was not only hard to derive, but it would also be difficult to implement in any official setting.

Our intention was to create a model that captured our population and thus could be used throughout the country given that we used data that took the average values of our features within each state in the US over a period of eight years. The limitations we find in our dataset, and the gaps of information within our datasets, however, do not allow us to conclude that our findings will completely reflect this population, especially given the lack of socioeconomic variables in our dataset.

Numerous studies on the Environment Health Perspectives (e.g. Estimating Causal Effects of Local Air Pollution on Daily Deaths: Effect of Low Levels, which we used for insight on instrumental variables associated with PM2.5) have suggested an association of long-term exposure to particulate matter (PM2.5) with mortality levels of various diseases.

Although our findings from OLS on PM2.5 and CVD-related mortality levels were not as successful, our research brought to our attention that there are strong cases that suggest that it is imperative to continue to research pollution levels caused by human production and develop laws that will prevent negative externalities that will harm the public.

Our group merged the CDC's U.S. Chronic Disease Indicators dataset with the EPA's daily summary data on the daily particulate data on PM2.5 mass and NO2 levels from 2010 to 2018. In addition, we combined the medicare expenditure from the Centers for Medicare (CMS), population sex proportions from the Kaiser Family Foundation (KFF), as well as diabetes, obesity prevalence, and tobacco consumption rates from the CDC's surveillance systems. One such benefit of combining data sources was that we were able to introduce an additional variable we could control within our model, but combining different sources consequently required the aggregation of many entries. One such consequence, for example, was in employing functions to determine annual averages for the datasets involving daily measures, such as PM2.5 levels, and re-inputting such calculations into the primary data frame.

The CDC's Annual State-Level US Chronic Disease Indicator dataset reported many different rates on an aggregate level. Since the data was not reported on an individual level, we believe a lot of the nuances in the risk factors were lost and difficult to account for in our analysis. Additionally, the mortality rates reported did not provide overlap in gender, age, and race, making it difficult to account for the interaction between these features.

The most interesting part of our work was finding that there continue to be uncaptured biases within government-sponsored datasets. There are assumptions that we as a society are aware of (such as the discrimination black people and other POCs suffer in terms of health care) that have yet to be fully explored and accounted for even in official datasets from the CDC and beyond. There will inevitably be implicit biases found within the development of any model, but future renditions of such studies could benefit from current findings. For instance, if we instead had data on individual patients in impoverished areas versus more affluent areas in a more controlled environmental setting, these biases may become more apparent in future studies involving the complex discussion involving disparities in race and health.

Work Cited

Bridges, K. M. (2018). Implicit bias and racial disparities in health care. Human Rights, 43(3), 19.

Crimmins, Eileen M., et al. Race/Ethnicity, Socioeconomic Status, and Health.

Lin, S. S., and J. L. Kelsey. "Use of Race and Ethnicity in Epidemiologic Research: Concepts, Methodological Issues, and Suggestions for Research." Epidemiologic Reviews, vol. 22, no. 2, 2000, pp. 187–202.,

Khera R, Valero-Elizondo J, Nasir K (October 2020). "Financial Toxicity in Atherosclerotic Cardiovascular Disease in the United States: Current State and Future Directions". Journal of the American Heart Association.

Mikkola, Tomi S et al. “Sex differences in age-related cardiovascular mortality.” PloS one vol. 8,5 e63347. 20 May. 2013, doi:10.1371/journal.pone.0063347

Schwartz, Joel, et al. “Estimating Causal Effects of Local Air Pollution on Daily Deaths: Effect of Low Levels.” *Environmental Health Perspectives*, National Institute of Environmental Health Sciences, Jan. 2017, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5226700/.

Sheppard, Lianne et al. “Confounding and exposure measurement error in air pollution epidemiology.” Air quality, atmosphere, & health vol. 5,2 (2012): 203-216. doi:10.1007/s11869-011-0140-9