Examine Whether Short-term Exposures to Air Pollution Will Affect Covid-19 Morality in Illinois State

AI and Development

December 13, 2022

Moeko Kondo, Sonali Subbu Rathinam and Miho Takahashi

1. Introduction

The World Health Organization defines air pollution as the 'contamination of indoor or outdoor environment by any chemical, physical, or biological agent that modifies the natural characteristics of the atmosphere' [I]. Exposure to air pollution in the long and short term is detrimental to human health, particularly to children or older age individuals, and those with pre-existing respiratory conditions. Even short time exposure to air pollutants can result in illnesses such as bronchitis and pneumonia. They also cause irritation to the nose, eyes, throat and skin, and cause headaches, dizziness and nausea [II]. Given that short term exposure exacerbates respiratory conditions, this project aims at examining whether short term exposure to air pollution has any effect on Covid-19 mortality. Since 2020, Covid-19 has been one the leading factors of death in the US [III], and mortality related risk factors of Covid-19 include (but not limited to) the aforementioned respiratory conditions that are also caused by short term exposure to air pollutants. Specifically, this project examines the effect of $PM_{2.5}$ and O_3 pollutants on Covid-19 mortality in Cook County in Illinois state from January 2020 to February 2021. This is done by implementing a Naive Bayes model and a Logistic Regression model on Covid-19 mortality and air pollutant data for Cook County.

The US Environmental Protection Agency (EPA) identified $PM_{2.5}$ and O_3 as two of the six air pollutants, hence they are used in this study. $PM_{2.5}$ refers to particulate matter (or tiny droplets in the air) that are less than 2.5 microns in aerodynamic diameter. Various scientific studies have shown that increased $PM_{2.5}$ exposure affects lung functioning, and escalates hospital admissions and deaths for those with underlying respiratory and cardiovascular conditions. $PM_{2.5}$ is produced by both outdoor sources such as automobile exhaust and fossil-fuel burning, and from indoor

sources such as tobacco smoke, cooking (frying and sautéing) and burning of candles [IV]. O_3 refers to ozone, and even low levels of ozone in the air can create respiratory problems, especially for those with chronic respiratory conditions, as its inhalation damages the lungs. O_3 is mostly produced by outdoor sources such as emissions from vehicles and factories [V]. Therefore, this study explores the relationship between the aforementioned air pollutants on Covid-19 mortality, and examines whether there are significant effects of the same. Identifying these relationships are crucial, as it helps in understanding factors that increase chances of hospitalizations and deaths due to Covid-19 in a region. This will help in implementing appropriate policy for emergency respiratory health services, air pollution and industrial sources that contribute to these air pollutants.

2. Data

The air pollutant data for this project has been obtained from supplemental sources of Kim et. al ‘s research paper titled “Association between Short-Term Exposure to Air Pollution and Covid-19 Mortality: A Population-Based Case Crossover Study Using Individual-Level Mortality Registry confirmed by Examiners” [VI]. This data for Cook County from January 2020 to February 2021 was collected from the EPA website. Since this includes the daily concentration levels of pollutants, this has been aggregated to three week averages, as it is more representative of their short-term impacts. This data was merged with Covid-19 mortality data obtained from Cook County’s official website. From this data, all missing values and deaths due to unnatural causes (such as accidents and suicide) were removed. Since the primary cause of death included causes that were not related to Covid-19 (e.g. cancer, urinary tract infections), the data was mapped such that all primary causes due to Covid-19 related complications were 1 and others 0.

This led to the creation of a binary dependent variable for the model. The final dataset has 6823 rows, and the unit of analysis is an individual in Cook County who died between January 2020 to February 2022. Table 1 provides further information on the dataset's variables.

Table 1: Data Dictionary Table for air pollutant and Covid-19 mortality dataset

Feature Description	Column Name in Data	Data Type	Specific Notes
Whether the Primary Cause of Death is Covid	Primary Cause Covid	Integer	Main Dependent Variable
The date the individual tested positive	Date	Date Type	Time Range: January 2020 to February 2021
Residential Area in Cook County	Residence City	String	After final merging, 13 unique residential areas are present
Ozone Concentration	O3	Float	This has been aggregated to 3-week average
$PM_{2.5}$ Concentration	PM25	Float	This has been aggregated to 3-week average
Logged Age	Age	Float	Logged age variable range is between (0, 4.8)
Whether the individual is a Latino	Latino	Integer	Race Indicator Variable
Whether the individual is an Asian	Asian	Integer	Race Indicator Variable
Whether the individual is an American Indian	Am. Indian	Integer	Race Indicator Variable
Whether the individual is Black	Black	Integer	Race Indicator Variable
Whether the individual is White	White	Integer	Race Indicator Variable
Whether the individual belongs to any other race not included in earlier columns	Other	Integer	Race Indicator Variable
Whether the individual is Male	Male	Integer	Gender Indicator Variable

From Figure 1, it can be inferred that there is a high prevalence of deaths in which the primary cause of death is not Covid-19. This leads to a class imbalance problem, and continuing with these results would lead to inaccurate results, as machine learning and artificial intelligence models will be heavily biased to the largely represented class (in this case, non-Covid-19 causes). In order to correct this, the random sampling method with logistic regression was performed. Essentially, oversampling and undersampling were considered to compensate for the imbalance already present in the data by introducing a bias to select more samples from one class. From Table 2, it can be inferred that the oversampling technique has slightly higher

accuracy, and hence the oversampled data was used to split into training and testing data eventually, Oversampled data selects random samples from the minority class (in this case, primary cause as Covid-19) with replacement, and supplements the data with multiple copies of these instances [VII].

On exploratory data analysis, it was found that there is a strong right skew in the distribution of the Age variable in the dataset. Hence, this variable was logged so that it is normally distributed, which ensures that variables do not bias results in regression models.

Figure 1: Distribution of Dependent Variable Primary Cause Covid-19

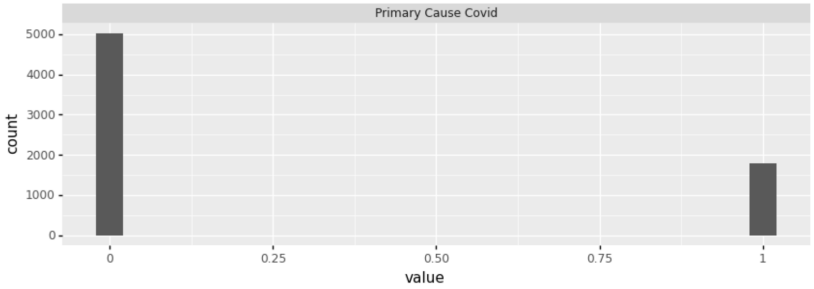
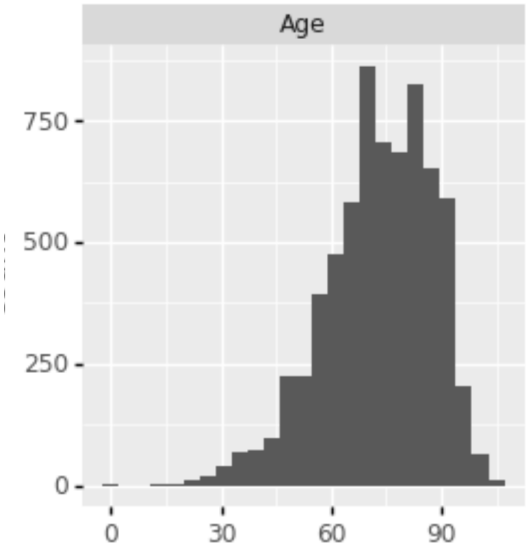


Table 2 : Accuracy Measures with Random Sampling Techniques

Class Balance Technique	Accuracy
Oversampling	54.5 %
Under sampling	48.4 %

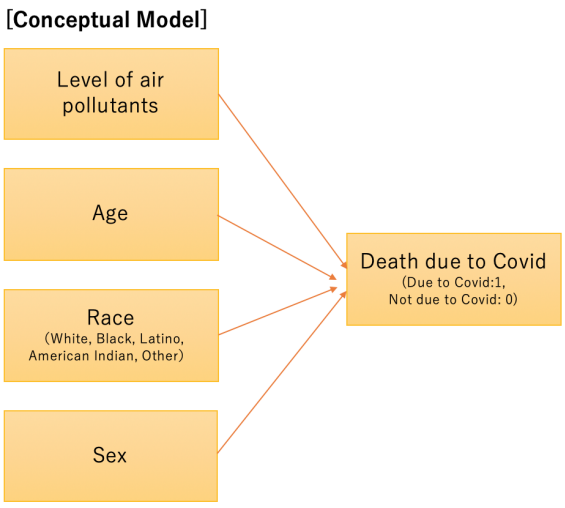
Figure 2 : Distribution of Age Variable before logging



3. Methods

We employed two methods, the Bayesian Network and logistic regression. Firstly, we used the Bayesian Network to graphically analyze the association between concentrations of air pollutants, age, sex, and race/ethnicity with Covid-19 mortality. We identified the variables and modeled the relationship based on knowledge from the research in this field. In this study, we modeled Naive Bayes, predicting the probability of Covid-19 mortality to examine whether those expected factors affect the Covid-19 mortality. We used Netica software to implement this analysis.

Secondly, to examine the association between concentrations of air pollutants, age, sex, and race/ethnicity with Covid-19 mortality, we also conducted a logistic regression analysis. We chose this method since the dependent variable is binary; whether the deaths are caused by Covid-19 (Primary cause is Covid-19; identifying this status as 1 in the model) or not (Primary cause is not Covid-19; in the model, 0). We used Stata software to implement this analysis.



4. Results

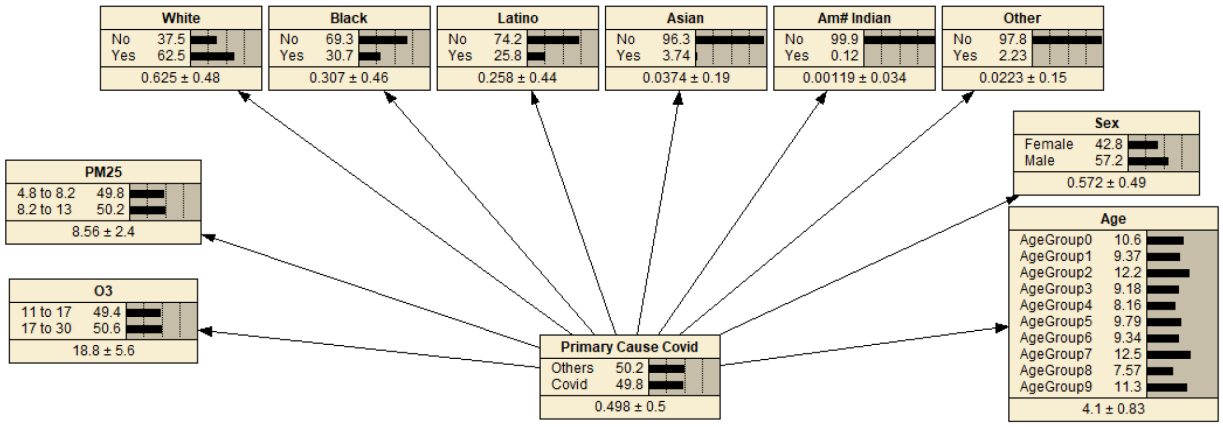
During the period we examined, 21-day averages of concentrations of air pollutants for every location based on the date of infection were 8.08μg/m3 for $PM_{2.5}$ and 17.36μg/m3 for O3. That is, all days were below the National Ambient Air Quality Standard (NAAQS); 35μg/m3 for 24 hours and 35.7μg/m3 for 8 hours, respectively [VIII, IX].

1) Bayesian Network

   a) Naive Bayes Model

The figure 3 represents our Naive Bayes model. We assumed that concentrations of air pollutants, age, sex, and race/ethnicity are associated with Covid-19 mortality. Our model is grounded in the knowledge of the aforementioned literature of the association between exposure to air pollution and Covid-19 mortality, and the other statistical reports of Covid-19 mortality by the Centers for Disease Control and Prevention [X].

Figure 3: Naive Bayes Model

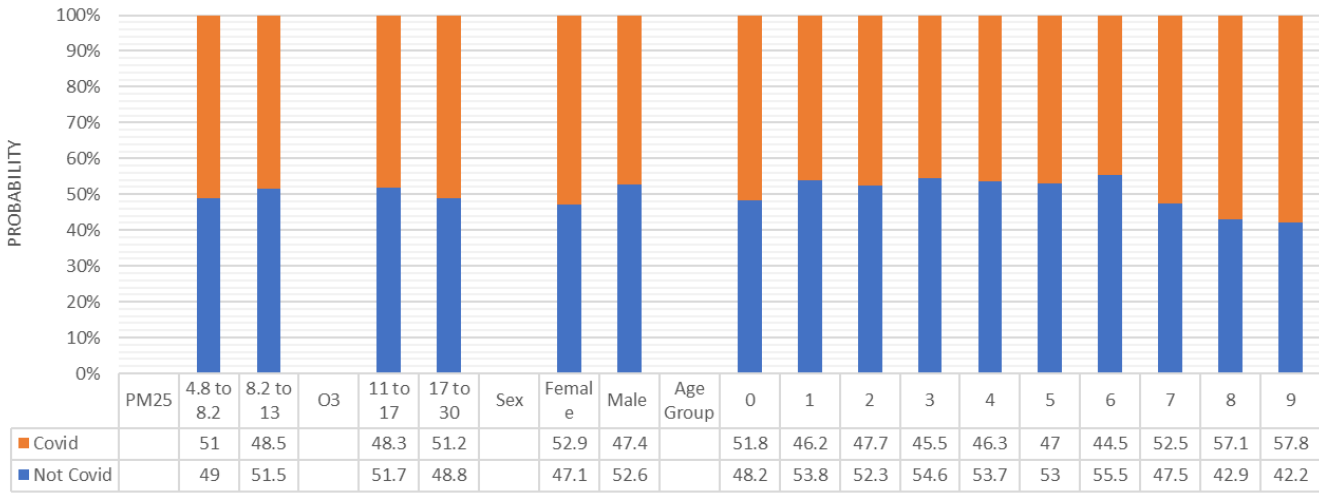


b) Results

The figure 4, 5 illustrates the conditional probabilities of death due to Covid-19. We observed each conditional probability by entering the particular state for each node.

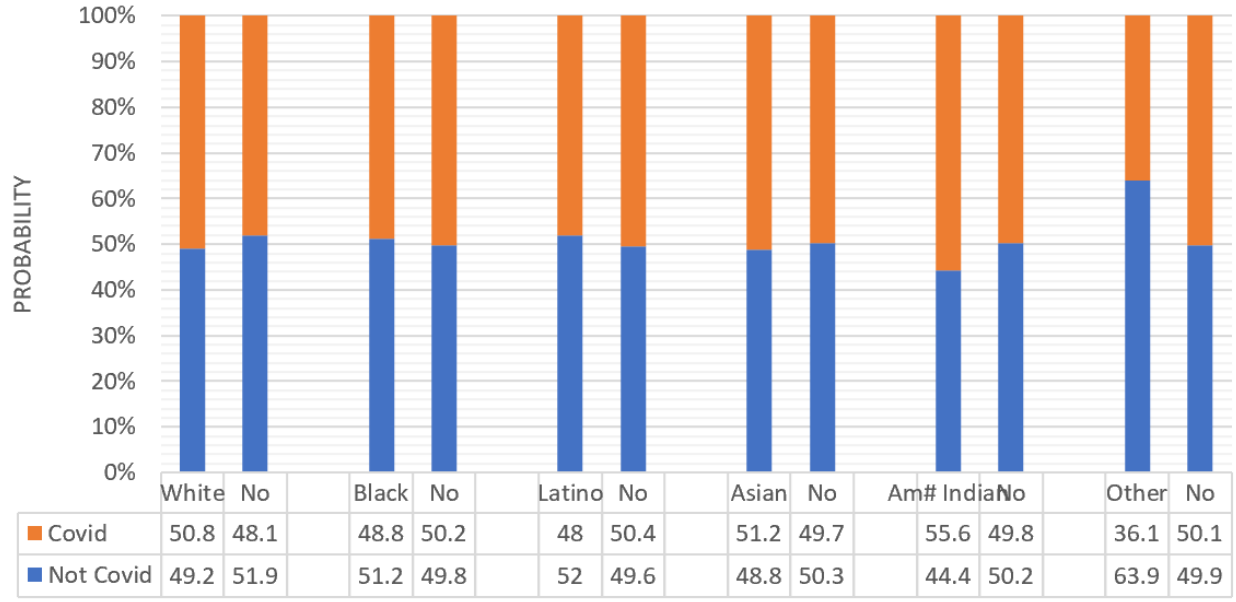
Our finding is that being female (sex) and O_3 increased the Covid-19 mortality by 5.5% and 2.9% respectively, while $PM_{2.5}$ decreased by 2.5%. Regarding race/ethnicity, being American Indian, white and Asian increased the Covid-19 mortality by 5.8%, 2.7% and 1.5% respectively, while being the other race, Latino, and black decreased the Covid-19 mortality by 14%, 2.4% and 1.4% respectively.

Figure 4: Conditional probabilities of cause of death given the particular state for air pollutants (PM2.5, O3), Sex and Age groups



	PM25	4.8 to 8.2	8.2 to 13	O3	11 to 17	17 to 30	Sex	Female	Male	Age Group	0	1	2	3	4	5	6	7	8	9
Covid		51	48.5		48.3	51.2		52.9	47.4		51.8	46.2	47.7	45.5	46.3	47	44.5	52.5	57.1	57.8
Not Covid		49	51.5		51.7	48.8		47.1	52.6		48.2	53.8	52.3	54.6	53.7	53	55.5	47.5	42.9	42.2

Figure 5: Conditional probabilities of cause of death given the particular state for race/ethnicity



	White	No		Black	No		Latino	No		Asian	No		Am# Indian	No		Other	No
Covid	50.8	48.1		48.8	50.2		48	50.4		51.2	49.7		55.6	49.8		36.1	50.1
Not Covid	49.2	51.9		51.2	49.8		52	49.6		48.8	50.3		44.4	50.2		63.9	49.9

Next, we conducted a sensitive analysis. The table 3 shows the sensitivity findings of this model. These sensitivity findings indicate how much the likelihood of death due to Covid-19 would be influenced by a single node state in the model. In this sensitivity analysis, the variable which influenced on Covid-19 mortality the most is age, followed by sex, being the other race, O_3 being

white, $PM_{2.5}$ and being Latino. These variables are more informative conditions in this model to predict the Covid-19 mortality correctly. On the other hand, the variable that had the lowest influence is being American Indian, followed by being Asian, black. The variance reduction score of these three variables are comparably low (under 0.0001); in other words, those variables had little association with Covid-19 mortality.

Table 3: Results of sensitivity analysis for Naive Bayes model

Sensitivity of 'Primary Cause Covid' to a finding at another node:

Node	Variance Reduction	Percent	Mutual Info	Percent	Variance of Beliefs
Primary Cause Covid	0.25	100	0.99998	100	0.2499943
Age	0.002057	0.823	0.00595	0.595	0.0020565
Male	0.0007461	0.298	0.00215	0.215	0.0007461
Other	0.000427	0.171	0.00125	0.125	0.0004270
O3	0.0002087	0.0835	0.00060	0.0602	0.0002087
White	0.0001702	0.0681	0.00049	0.0491	0.0001702
PM25	0.0001576	0.063	0.00045	0.0455	0.0001576
Latino	0.0001047	0.0419	0.00030	0.0302	0.0001047
Black	3.79e-05	0.0152	0.00011	0.0109	0.0000379
Asian	8.453e-06	0.00338	0.00002	0.00244	0.0000085
Am# Indian	4.016e-06	0.00161	0.00001	0.00116	0.0000040

To summarize, firstly, we could not find the association between $PM_{2.5}$ and Covid-19 mortality as we hypothesized, and a relationship between age and Covid-19 mortality since the mortality did not increase consistently along with age. Secondly, being female (sex) and the other race had relatively strong associations with Covid-19 mortality, increasing and decreasing the probability of death due to Covid-19 respectively. In addition to that, O_3 and being white had an association with Covid-19 mortality, contributing to increase of Covid-19 mortality. At the same time, being Latino also had an association with Covid-19 mortality, but decreasing the probability. Lastly, sensitivities for being black, Asian and American Indian are quite low, and there were little relationships between Covid-19 mortality.

At last, we conducted a test with 25% of our dataset in this model, obtaining an error rate of 45.87%, and the area under ROC (AUC) of 0.5505. This AUC score indicates that our model could classify the Covid-19 mortality slightly more accurately than random guessing of AUC 0.5.

2) Logistic Regression Model

Table 4 presents the results of logistic regressions. In this regression, $PM_{2.5}$, age, being American Indian, and other race do not have a statistically significant relationship with Covid-19 mortality (The p-values of these variables are above all conventional levels of significance: $p<0.1$, $p<0.05$, or $p<0.01$). The result for O_3 shows an odds ratio of 1.037(95% confidence interval (CI): 1.03, 1.05). There is a 95% probability that every unit increase in O_3 concentration increases the likelihood of a person dying from Covid-19, having all others equal. This finding is significant at $p< 0.01$ level. The odds ratio for males is 0.79 (CI: 0.73, 0.86), meaning males have a negative and significant correlation with Covid-19 mortality. That is, females are more likely to die because of Covid-19 than males, and this finding is significant at $p< 0.01$ level. For race/ethnicity, the odds ratio for Latino, Asian, Black, and White are 0.91 (CI: 0.82, 1.00), 1.75 (CI: 1.06, 2.90), 1.50 (CI: 0.94, 2.40), and 1.68 (CI: 1.06, 2.68), respectively. Being Asian, Black or White contributes to the increase of the probability of Covid-19 mortality, while being Latino negatively affects the probability of death due to Covid-19. The findings for Asians and White are significant at $p< 0.05$ level, and those for Latinos and Black are at $p< 0.1$ level. Note that since the 95% confidence interval of Black straddles 1, there is more than a 5 % possibility that being Black contributes to the increase of the probability of death caused by not Covid-19. Comparing these odds ratios, the specific race or gender contributes more to the probability of death by Covid-19 than the concentration levels of O_3.

Table 4. Results of a logistic regression: testing the relationship between the concentration of air pollutants ($PM_{2.5}$ and O_3) and other variables, with whether Covid-19 caused death as the dependent variable.

```
. logit PrimaryCauseCovid PM25 O3 Age Latino Male AmIndian Asian Black Other White, or

Iteration 0:   log likelihood = -6985.5373
Iteration 1:   log likelihood = -6931.2218
Iteration 2:   log likelihood = -6931.2035
Iteration 3:   log likelihood = -6931.2035

Logistic regression                                     Number of obs = 10,078
                                                        LR chi2(10)   = 108.67
                                                        Prob > chi2   = 0.0000
Log likelihood = -6931.2035                             Pseudo R2     = 0.0078
```

PrimaryCauseCovid	Odds ratio	Std. err.	z	P>|z|	[95% conf.	interval]
PM25	.992571	.0172283	-0.43	0.667	.9593722	1.026919
O3	1.03672	.005863	6.38	0.000	1.025292	1.048275
Age	.993351	.0878538	-0.08	0.940	.8352585	1.181366
Latino	.9053746	.0477505	-1.88	0.059	.8164601	1.003972
Male	.7915716	.0328542	-5.63	0.000	.729728	.8586563
AmIndian	2.303623	1.586662	1.21	0.226	.5972156	8.885703
Asian	1.751175	.4514461	2.17	0.030	1.056557	2.902458
Black	1.498575	.3589699	1.69	0.091	.9370892	2.396492
Other	.8863875	.2417534	-0.44	0.658	.519358	1.512796
White	1.683696	.3985871	2.20	0.028	1.058658	2.677759
_cons	.4290134	.2099415	-1.73	0.084	.1644099	1.119474

Note: **_cons** estimates baseline odds.

5. Conclusion

Our findings showed that O_3 concentrations were associated with Covid-19 mortality, but the results for $PM_{2.5}$ were not statistically significant. From these results, it can be inferred that reducing O_3 concentrations may reduce the risk of Covid-19 mortality. We also found that gender and specific race/ethnicity were more strongly associated with Covid-19 mortality than O_3 concentrations. During the study period, air quality in Cook County was consistently below the ambient air quality standards (NAAQS), suggesting that short-term exposure to pollutants over three weeks had a relatively small effect on Covid-19 mortality.

There are also several limitations to this study. First, we classified whether deaths were due to Covid-19 or not based on the primary cause of death on the death certificates of those infected with Covid-19 and died. However, the criteria for determining the primary cause of death on the

death certificate are unknown; for example, it is possible that people who died with pneumonia could have had severe pneumonia due to Covid-19 (i.e., we might be able to classify Covid-19 as the primary cause of death for these people). We assume that we can examine the relationship between air pollution concentrations and other variables with Covid-19 mortality more accurately by comparing those infected with Covid-19 but did not die with those who did die. Still, we were unable to obtain such a data set. Second, as noted above, Cook county air quality during the study period was below the NAAQS. Therefore, it is possible that the study did not accurately determine the magnitude of the impact of short-term exposure to pollutants on Covid-19 mortality. Furthermore, following previous research, we set the short-term exposure period to 3 weeks, but extending that period could yield different results. It would be desirable to conduct similar studies in the future with adjusted exposure periods in areas where air quality exceeds the NAAQS. Third, our analysis could not examine the effects of an individual's chronic diseases or disabilities, economic status, or other health behaviors (e.g., smoking) because of data availability. These variables may affect mortality of Covid-19.

6. Reference

I : World Health Organization. (n.d.). *Air Pollution*. World Health Organization. Retrieved December 11, 2022, from https://www.who.int/health-topics/air-pollution

II : *Air Pollution*. National Geographic Society. (n.d.). Retrieved December 11, 2022, from https://education.nationalgeographic.org/resource/air-pollution

III : Centers for Disease Control and Prevention. (2022, April 22). *Covid-19 was third leading cause of death in U.S.* Centers for Disease Control and Prevention. Retrieved December 11, 2022, from https://www.cdc.gov/media/releases/2022/s0422-third-leading-cause.html

IV : *Department of Health*. Fine Particles (PM 2.5) Questions and Answers. (n.d.). Retrieved December 11, 2022, from

https://www.health.ny.gov/environmental/indoors/air/pmq_a.htm#:~:text=Particles%20in%20the%20PM2.5%20size%20range%20are%20able%20to,nose%20and%20shortness%20of%20breath

V : *California Air Resources Board*. Ozone & Health | California Air Resources Board. (n.d.). Retrieved December 11, 2022, from https://ww2.arb.ca.gov/resources/ozone-and-health#:~:text=Where%20does%20ozone%20come%20from,paints%2C%20and%20many%20other%20sources.

VI : Kim H, Samet JM, Bell ML. Association between Short-Term Exposure to Air Pollution and COVID-19 Mortality: A Population-Based Case-Crossover Study Using Individual-Level Mortality Registry Confirmed by Medical Examiners. *Environ Health Perspect*. 2022 Nov;130(11):117006.doi:10.1289/EHP10836.

VII: Pykes, K. (2020, September 10). *Oversampling and Undersampling*. Medium. Retrieved December 11, 2022, from https://towardsdatascience.com/oversampling-and-undersampling-5e2bbaf56dcf

VIII: *National Ambient Air Quality Standards (NAAQS) for PM.* (2022, July 19). United States Environmental Protection Agency. Retrieved December 11, 2022, from https://www.epa.gov/pm-pollution/national-ambient-air-quality-standards-naaqs-pm

IX: *Ozone National Ambient Air Quality Standards (NAAQS).* (2022, February 9). United States Environmental Protection Agency. Retrieved December 11, 2022, from https://www.epa.gov/ground-level-ozone-pollution/ozone-national-ambient-air-quality-standards-naaqs

X: *Risk of COVID-19-Related Mortality.* (2022, November 16). Centers for Disease Control and Prevention. Retrieved December 11, 2022, from https://www.cdc.gov/coronavirus/2019-ncov/science/data-review/risk.html