Effects of PM$_{2.5}$ and O$_3$ on Human Health at a Suburban Area of Beijing, China

Siyuan (Larry) Chen$^1$, Michelle Dai$^2$, Fengli Lin$^3$

$^1$William G. Enloe High School, Raleigh, NC, USA
$^2$North Carolina School of Science and Mathematics, Durham, NC, USA
$^3$Maternal & Child Health Care Center, Yunhe District, Cangzhou City, China

Email: larrychen37@gmail.com

Abstract

This work investigates the impact of air pollution on human health in China. First-hand data obtained at a suburban area of Beijing between January 2014 and June 2016 forms the basis for our study. It is hypothesized that there is a statistically significant relationship between lung cancer incidences and levels of air pollutants, in particular, PM$_{2.5}$ and O$_3$. Methods of analysis include correlation analysis, multiple regression, and trend analysis. Our results verify the hypothesis, and reveal that the high rates of lung cancer development in the studied period of 30 months can be interpreted by the high concentration of PM$_{2.5}$ and O$_3$ using the Poisson linear regression model. It is also discovered that PM$_{2.5}$ and O$_3$ have a strong negative correlation and they alternatingly serve as the primary contributor of high rates of lung cancer in cold and warm seasons. Also, men are found to have a higher rate of developing lung cancer than women, and middle-aged people are most likely to develop lung cancer compared to the elderly and youth.

Keywords

PM$_{2.5}$, O$_3$, Air Pollution, Lung Cancer

1. Introduction

Air pollution has been an inevitable side product in the process of industrialization and urbanization. The exposure to air pollutants, especially particulate matter (PM), causes the death of millions of people and the loss in billions of dollars [1] [2] [3] [4] [5]. Its impact on human health has received great attention in recent years [6]–[13]. China, as the largest developing country and the
second largest economic body in the world, has been undergoing rapid economic growth in the last 30 - 40 years. They also paid a big price for this development—the environmental pollution [14]. It was reported that in 2013, the annual mean PM$_{2.5}$ concentration in Beijing reached 100 micrograms per cubic meter, which is about 3 times the interim target-1 level of the World Health Organization (WHO) [13] [15] [16]. Beijing is by no means the only city in China that suffers from severe air pollution; in fact, more than 90% of Chinese cities have an annual mean PM$_{2.5}$ concentration that exceeds the above WHO standard, according to a recent record of the China Environmental Status Bulletin. This situation has raised serious concerns on the impact of air pollution to the people’s health in China [17] [18] [19].

Many elements contribute to air pollution, including total suspended particles (TSP), particulate matter (PM)—in particular PM$_{10}$ and PM$_{2.5}$ as measured by the particle diameter, ozone (O$_3$), Carbon monoxide (CO), Sulfur dioxide (SO$_2$), nitrogen dioxide (NO$_2$), etc. Among these many pollutants, fine particulate matter (PM$_{2.5}$) and ozone (O$_3$) are suggested as the two main harmful contaminants endangering human health, in particular promoting lung diseases [20].

Quantitative studies of the impact of air quality on human health have recently received increased attention. In the following, a brief review is given on some representative works. Li et al. [21] studied the effects of polycyclic aromatic hydrocarbons (PAH’s), in surface water, sediments, soils, and plants via air, water, and food pollution, on the occurrences of cancers. They used a variety of mathematical equations derived from experience or experiments to calculate correlations. Buonanno et al. [9] estimated lung cancer risk associated with airborne molecules of various elements. They utilized a modified risk-assessment scheme to estimate the contribution to cancer risks from sub-micron and super-micron molecules and proposed ways to improve air quality in Europe. Chen et al. [22] investigated the relationship between long-term exposure to PM$_{10}$, NO$_2$, and SO$_2$ and mortality of lung cancer in northern China. Hazard ratios were estimated using Cox proportional hazard models, adjusted for age, gender, and other factors. Correlation between PM$_{10}$ concentrations and lung cancer mortality was found, and associations varied across many categories. Dziubanek et al. [10] analyzed the impact of PM$_{10}$, benzo(a)pyrene (BaP), cadmium (Cd) and lead (Pb) on the lifespan of 3.5 million people living in the Silesia province in Poland. They applied Pearson linear regression to determine the relation between certain elements and lifespan, and multiple regression analysis to examine the effects of mixing multiple elements. Wu et al. [13] used an exposure-response function to quantitatively estimate the health effects attributed to PM$_{2.5}$ in China. Raaschou-Nielsen et al. [23] conducted a study on the individual elements that make up PM$_{10}$ and PM$_{10}$ particles to determine which elements have a greater correlation with lung disease incidence. Though there were elevated levels of correlation, no statistically significant results were found concerning the correlation between element amounts and hazard ratios for lung cancer. When the analyses were restricted to people who did not change residences during follow-up, statis-
tically significant results for PM$_{2.5}$ Cu, PM$_{10}$ K, and PM$_{2.5}$ S were revealed. The closest work to ours is Guo et al. [20], which studied the association between incidences of lung cancer and PM$_{2.5}$ and O$_3$ levels, separately, using two-year average data.

In summary, most existing quantitative studies either focused on pollutants different from what we consider, or focused on PM$_{2.5}$ only. Guo et al. [20] considered the effect of both PM$_{2.5}$ and O$_3$ on lung cancer incidence based on yearly-averaged data. However, the air quality varies significantly between seasons as shown in our data source, which was not captured in Guo’s study. Besides, the PM$_{2.5}$ and O$_3$ themselves usually have strong correlations with each other on a seasonal scale [24], which means they should be jointly considered when exploring their relationship with human health.

In this study, using the first-hand data obtained in a suburban area of Beijing during the period from January 2014 to June 2016, we investigate the relationship between lung cancer incidences and air pollution, specifically examining the effects of PM$_{2.5}$ and O$_3$. Our work has two distinctions compared to exiting works. Firstly, we use monthly data to discover the relationship between the levels of air pollutants and lung cancer incidences on a seasonal scale. Furthermore, we consider the correlation between PM$_{2.5}$ and O$_3$ and explore their joint effects on human health.

2. Materials & Methods

The data of lung cancer patients on a daily scale was obtained from the Center for Disease Control located in Yunhe District of Cangzhou City, a suburban area of Beijing, China (Figure 1). Over one million people live in the metro area, and a total of over eight million people live in the city of Changzhou. Patient information, including ages and genders, were included in the dataset. The concentrations of two main air pollutants, PM$_{2.5}$ and O$_3$, were obtained from the China Air Quality Online Monitoring and Analysis Platform (https://www.aqistudy.cn/historydata/about.php) on a monthly scale.

First, the raw daily data of patients were processed into a monthly form to match with the data of air pollutants. The patients were then categorized according to their genders and ages. Three age groups are considered in our study, namely, the youth (younger than 45 years old), the middle-aged (between 45 and 65 years old), and the elderly (older than 65 years old).

Inspired by existing literature, we intend to extend on preexisting research to evaluate the relationship between the two air pollutants, i.e. PM$_{2.5}$ and O$_3$, and lung cancer incidences on a monthly scale. Our hypothesis is that there is a statistically significant relationship between them, and we intend to find a meaningful regression model for this relationship. Given the nature of our data source, finding such a model was highly challenging. We started with the linear regression models between the lung cancer incidences ($Y$) and PM$_{2.5}$ ($X_1$) and O$_3$ ($X_2$) individually and jointly, gradually enriching the model by considering the
interactions ($X_1^1 \times X_2^1$), quadratic terms ($X_1^2$ and $X_2^2$), and polynomial fitting with 10 terms up to cubic level predictors ($X_1^3$ and $X_2^3$). Unfortunately, all these trials were unsuccessful, yielding high p-values between 0.5 and 0.9. It was further observed that fitting the data with higher order polynomials actually led to worse performance, which indicates that the relationship we intend to characterize may be nonlinear. Furthermore, we realized that the response variable $Y$ here stands for the number of people (count data), which only takes integer values and violates the classic assumption of normal distribution for linear regression. With these considerations, we then further explored the Generalized Linear Model (GLM) for a better fitting. In particular, we assume a Poisson distribution for the response variable and transform the response variable to the logarithm domain. It turns out that this model resulted in a statistically significant p-value when the effect of PM$_{2.5}$ ($X_1$) and O$_3$ ($X_2$) are jointly considered.

We plotted the accumulation curves of the number of male and female patients in the 30 months period studied. The increasing slopes of the accumulation curves were quantitatively compared in different seasons to examine their variations with genders and seasons.

### 3. Results

In our collected raw data, there are 3459 lung cancer patients in total, in which 41% are females. The proportion of female patients has risen from 38.4% to 43.6% from 2014 to 2016. The proportions of patients in the age categories of youth, middle-aged, and elderly are about 6%, 58%, and 36%, respectively.

We processed the original data in Excel and conducted statistical analysis us-

---

**Figure 1.** The city (Cangzhou) of our study in China.
ing MATLAB. As a first step, we examined the correlation between levels of PM$_{2.5}$ and O$_3$. Correlation analysis reveals that the Pearson correlation coefficient of the two air pollutants is close to $-0.7$ with a $p$-value on the order of $10^{-5}$, which verifies that there is a strong negative correlation between the levels of PM$_{2.5}$ and O$_3$, as indicated in [24]. This finding sets the foundation for our following study.

We further explored various regression models for the relationship between the lung cancer incidences ($Y$) and PM$_{2.5}$ ($X_1$) and O$_3$ ($X_2$), and finally determined that Generalized Linear Model with Poisson distribution and the following logarithm link function gives the best fit with statistical significance:

$$\log Y \sim 1 + X_1 + X_2$$

(1)

The results are shown below in Table 1.

We would like to note that in contrast to the results in [20], applying the same regression model to PM$_{2.5}$ and O$_3$ separately does not yield models with statistical significance (with $p$-values of 0.737 and 0.0724, respectively). This point will be further discussed in Section 4.

We further applied the same statistical model to subgroups of patients by gender and age. The corresponding results are presented below in Tables 2-6.

**Table 1.** Generalized linear model: Total number of patients vs. PM$_{2.5}$ and O$_3$.

<table>
<thead>
<tr>
<th>Estimate (SE)</th>
<th>t-Stat</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>4.4512 (0.11652)</td>
<td>38.202</td>
</tr>
<tr>
<td>$X_1$</td>
<td>0.0019936 (0.00091894)</td>
<td>2.1694</td>
</tr>
<tr>
<td>$X_2$</td>
<td>0.0015566 (0.0005565)</td>
<td>2.7971</td>
</tr>
</tbody>
</table>

**Table 2.** Generalized linear model: Total number of female patients vs. PM$_{2.5}$ and O$_3$.

<table>
<thead>
<tr>
<th>Estimate (SE)</th>
<th>t-Stat</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>3.876 (0.18026)</td>
<td>21.502</td>
</tr>
<tr>
<td>$X_1$</td>
<td>-0.00063228 (0.0014372)</td>
<td>-0.43992</td>
</tr>
<tr>
<td>$X_2$</td>
<td>0.00034342 (0.00086791)</td>
<td>0.39568</td>
</tr>
</tbody>
</table>

**Table 3.** Generalized linear model: Total number of male patients vs. PM$_{2.5}$ and O$_3$.

<table>
<thead>
<tr>
<th>Estimate (SE)</th>
<th>t-Stat</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>3.7019 (0.15278)</td>
<td>24.231</td>
</tr>
<tr>
<td>$X_1$</td>
<td>0.0038138 (0.001196)</td>
<td>3.1889</td>
</tr>
<tr>
<td>$X_2$</td>
<td>0.0024066 (0.00072558)</td>
<td>3.3169</td>
</tr>
</tbody>
</table>

**Table 4.** Generalized linear model: Total number of youth patients vs. PM$_{2.5}$ and O$_3$.

<table>
<thead>
<tr>
<th>Estimate (SE)</th>
<th>t-Stat</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>2.1481 (0.464)</td>
<td>4.6295</td>
</tr>
<tr>
<td>$X_1$</td>
<td>0.00012877 (0.0036462)</td>
<td>0.035315</td>
</tr>
<tr>
<td>$X_2$</td>
<td>-0.0020621 (0.0022785)</td>
<td>-0.90504</td>
</tr>
</tbody>
</table>
Table 5. Generalized linear model: Total number of middle aged patients vs. PM$_{2.5}$ and O$_3$.

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>SE</th>
<th>t-Stat</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>3.817</td>
<td>0.15375</td>
<td>24.826</td>
<td>4.6959e-136</td>
</tr>
<tr>
<td>$X_1$</td>
<td>0.0029713</td>
<td>0.0012061</td>
<td>2.4637</td>
<td>0.013753</td>
</tr>
<tr>
<td>$X_2$</td>
<td>0.0016649</td>
<td>0.00073403</td>
<td>2.2681</td>
<td>0.023322</td>
</tr>
</tbody>
</table>

Table 6. Generalized linear model: Total number of elderly patients vs. PM$_{2.5}$ and O$_3$.

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>SE</th>
<th>t-Stat</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>3.4266</td>
<td>0.19513</td>
<td>17.56</td>
<td>4.9566e-69</td>
</tr>
<tr>
<td>$X_1$</td>
<td>0.00081009</td>
<td>0.0015554</td>
<td>0.52083</td>
<td>0.60248</td>
</tr>
<tr>
<td>$X_2$</td>
<td>0.002387</td>
<td>0.0009261</td>
<td>2.5775</td>
<td>0.0099523</td>
</tr>
</tbody>
</table>

We observed that statistical significance was achieved for the male group and the middle-aged group, partially for the elderly group, but not for others. A further discussion will be given in the next section.

The above statistical analysis indicates that PM$_{2.5}$ and O$_3$ jointly make significant contributions to lung cancer instances, while they have strong negative correlation. This motivates us to further explore their individual roles in different seasons.

Figure 2 shows how the accumulated number of lung cancer patients varies within the 30 months period studied. Based on Figure 2, the accumulated lines for males (blue line in Figure 2) are remarkably steeper than the accumulated lines for females (red line in Figure 2); this means in general, males are more sensitive to the persistent air pollution than females, although data statistics indicate that the proportions of female patients have risen from 38.4% to 43.6% from 2014 to 2016. Steep slopes of the accumulation curves occur in the summer and winter seasons in 2014 and from the summer of 2015 to the spring of 2016. To further investigate their variations with seasons and discover the dominant air pollutant in different seasons, we also examine below the accumulated numbers of lung cancer patients vs. accumulated levels of PM$_{2.5}$ and O$_3$ in Figure 3(a) and Figure 3(b), respectively.

Aside from the similar phenomena between genders, Figure 3 highlights various windows in summers and winters with rapidly increasing numbers of patients. For example, Figure 3(b) indicates that lung cancer incidents increased rapidly in the winter seasons (October-January). Based on Figure 4, the slopes of the linear regression approximately doubled and tripled in the periods of Oct., 2014-Jan. 2015 and Oct. 2015-Jan. 2016, respectively. It is interesting to realize that the O$_3$ levels were relatively low during the above winter seasons, which accounts for the rapid increase in the slopes of the accumulation curve; however, the PM$_{2.5}$ levels were relatively high during these periods in north China due to the coal-powered heating supplies.
Figure 2. Accumulated number of lung cancer patients.

Figure 3. Accumulated number of lung cancer patients vs. (a) accumulated PM$_{2.5}$; (b) accumulated O$_3$. 
In contrast, Figure 3(a) indicates two summer periods with steep slopes when the O₃ levels were relatively high and PM₂.₅ levels were relatively low. Combined with the negative correlation between PM₂.₅ and O₃ found in the previous section, our results strongly suggest that the two main air pollutants, i.e. PM₂.₅ and O₃, may alternately play a dominant role in the winter seasons and summer seasons, respectively.

4. Discussion

Our results provide new insights on the impacts of air pollution on human health. It is already well known that many air pollutants, such as PM₂.₅ and O₃, have adverse effects on human health, especially on the respiratory system of human beings. However, most previous studies analyzed data on a yearly scale and have not looked into the alternating dominant role of PM₂.₅ and O₃ in winters and summers, which have a negative correlation on a seasonal scale. Our data is set on a monthly scale from 2014 to 2016, therefore our analysis can provide more insights for relevant studies in the same area.

As a comparison, Guo et al. [20] studied a similar problem using data of a much larger scale in both time and spatial dispersion. Data of lung cancer incidence from 75 communities covering a wide range of China from 1990 to 2009 were used, and statistically significant associations were found between lung cancer incidence and PM₂.₅, and between lung cancer incidence and O₃, separately. There are several differences between this work and ours. 1) Our data is limited in scope, so we could not find statistically significant associations between lung cancer incidence and PM₂.₅/O₃, separately. This may not necessarily indicate flaws, but rather suggests that the data source should be jointly considered when interpreting different results. 2) We successfully identified the association of lung cancer incidence with PM₂.₅ and O₃ jointly, which was not studied in [20]. Both results confirm the linear effects of PM₂.₅ and O₃ on (the logarithm transform) of lung cancer incidence. 3) As Guo et al. [20] used data on a yearly
scale, they could not demonstrate changes on a season scale as we did, nor did they identify the different roles PM$_{2.5}$ and O$_3$ play in different seasons.

Compared to PM$_{2.5}$, O$_3$ has more complicated effects on the respiratory system of human beings. As a powerful oxidant, ozone may damage mucous and respiratory tissues in human beings. Therefore, although the ozone layer (a portion of the stratosphere with a high concentration of ozone at about two to eight ppm) is beneficial in preventing damaging ultraviolet light from reaching the Earth’s surface, ozone is a potential respiratory hazard and pollutant near the ground level. Besides, ozone itself is a production of the photochemical reaction (by strong ultraviolet light) of nitrogen oxide (NOx) and volatile organic compounds (VOCs) coming from pollutant sources such as industrial and automobile exhaust. Therefore, the high values of the O$_3$ level, present in the summer seasons, also indicate a relatively high level of primary pollution from industries and vehicles during those seasons.

The negative relationship between PM$_{2.5}$ and O$_3$ could be because the PM$_{2.5}$ is highly related to heating supplies used in winter seasons in China, while the production of O$_3$ requires strong ultraviolet light, which is more present in the summer seasons.

There are some limitations in our study. As mentioned earlier, our data is limited in scope. In our statistical analysis, we could not verify statistically significant association of lung cancer incidence with PM$_{2.5}$ and O$_3$ for the female group and the youth group. The latter could be attributed to the small data sample (accounting for only 6% of the total). Also, we did not control for other factors that contribute to lung cancers, such as smoking and different lifestyles.

It is interesting to note that patients in the age group 45 - 65 have the highest proportion of 58%. Many assume that the elderly would have the highest proportion of lung cancer rates, but our data show that the elderly make up 36% of the total data. This could be explained by recent reports of a trend that cancer patients are getting younger. Not only do young and middle-aged people have a faster metabolism and cell activity, nowadays they also tend to live an unhealthy lifestyle of staying up late, excessive alcohol drinking, etc. Another possible reason is the excessive anxiety and pressure, which the middle-aged are experiencing with the rapid growth of the Chinese economy. In addition, our analysis indicates that lung cancer rates for the youth are insignificantly low. It is easy to understand, as the carcinogenesis of lung cells need accumulated stimulations from the environment.

In addition, our sampling pool indicates that the proportion of female patients of lung cancers has risen from 38.4% to 43.6% from 2014 to 2016 (42.9% in the year of 2015). It might be related to modern cooking habits in China—it has been reported that more than half of female patients who got lung cancers have been exposed to kitchen fumes over a long period of time. In addition, there are more and more female smokers in China; The French Association of Pulmonary Surgeons believes that females are more vulnerable to nicotine than males do, which increases lung cancer risks.
5. Conclusions and Future Work

Through this study, it is found that the PM$_{2.5}$ and the O$_3$ have significant adverse effects on the respiratory system of human beings. Methods of analysis include correlation analysis, multiple regression, and trend analysis. In contrast to previous studies, the present research is set on a monthly scale, which can provide more insights for relevant studies in the same area. Our analysis indicates PM$_{2.5}$ and O$_3$ have a negative correlation on a seasonal scale. PM$_{2.5}$ and O$_3$ also alternatingly play a dominant role in winters and summers, respectively. The above conclusions are drawn based on our current data collected in a suburban area of Beijing.

This study also finds that men have a higher rate of lung cancer development than women, and middle-aged people (45 - 65) are most likely to develop lung cancer compared to the elderly (>65) and youth (<45). The trend of cancer patients becoming younger is attributed, at least partially, to the unhealthy habits and lifestyles of the modern society.

Much work remains to be done. In our study, we classified all lung cancers as one dependent variable; each specific lung cancer could be singled out, and the effects of pollution on each specific cancer could be further investigated.

We also plan to collect more data, perhaps from other areas as well, to further examine the associations of lung cancer incidence with PM$_{2.5}$ and O$_3$ for the subgroups that we could not give a definite answer for in this study.

Then, we plan to further extend the scope of our study. Are there other pollutants besides PM$_{2.5}$ and O$_3$ that have significant impacts on the respiratory system of human beings? What would be their interactions with PM$_{2.5}$ and O$_3$, and how would they jointly influence human health?

Finally, we plan to explore other important contributors for lung cancers that are not included in our current research, such as smoking. By removing the impacts of these factors, a better understanding of the effects of air pollution to human health may be achieved.

Acknowledgements

We wish to express our deep gratitude to the Center for Disease Control located in Yunhe District of Cangzhou City for permission to access the lung cancer data.

References


