Dissecting the Risk Factors for Lung Cancer

Yuwei Cui
ywcu@umd.edu
Neuroscience and Cognitive Sciences
University of Maryland, College Park

Jin Sun
jinsun@cs.umd.edu
Department of Computer Science
University of Maryland, College Park

Introduction

Lung cancer is one of the most common cancer types in the United States, and has a tremendous impact on the public health. In the year of 2013, there are an estimated 228,190 new cases and 159,480 deaths according to the American Cancer Society [1]. Although epidemiologic researches have quickly identified the tight association between cigarette smoking and lung cancer [2], there is also a persistent public concern that air pollution is causing lung cancer [3]. It is a known fact that carcinogens are continuously released into the air from motor vehicles, power plants, and other industrial sources. However, the evidence supporting an association between air pollution and lung cancer is mixed. Earlier researches using case-control and cohort studies did not support the association, while recent studies report evidences that suggest a genuine association [4].

One difficulty of establishing the link between air pollution and lung cancer incidence in the early studies is the lack of long-term trend data for outdoor levels of known air pollutant [3]. Since the 1990’s, the US Environmental Protection Agency (EPA) collects nationwide data on major air pollutant. On the other hand, the Centers for Disease Control (CDC) and the National Cancer Institute (NCI) will combine their registries and generate a set of official statistics on cancer incidence every year. In this report, we have utilized the state-of-the-art information visualization techniques to explore the potential risk factors for lung cancer. We find significant correlations between lung cancer incidence rate and air pollutant concentrations across the United States, as well as persistent geographical differences of lung cancer incidence rate. We also recognize an interesting gender difference in lung cancer susceptibility.

Methods

Data Source

Cancer Statistics Data
We retrieve the lung cancer incidence data from the United States Cancer Statistics (USCS) website, which is produced by the CDC and the NCI. This dataset contains state specific cancer incidence data from registries that have high-quality data for 1999-2009.
Air Pollution Data
We extract air pollution data from the Air Quality System (AQS), which is EPA’s repository of ambient air quality data. This dataset contains air pollution data collected by EPA, state, and local air pollution control agencies from thousands of monitor stations since 1993. We used the data on six major pollutants (Carbon Monoxide, Lead, PM2.5, PM10, Nitrogen Dioxide and Sulfur Dioxide) from 1999 to 2009, in parallel with the cancer statistics data.

Data Pre-process
We have filtered, aggregated, and re-organized the data before using the visualization software. For cancer statistics data, we only used entries with “lung cancer”. For the air pollution data, we calculated the average concentration for each pollutant within each state for each year. We then combine the cancer statistics data and the air pollution data into a single file, with each row lists the state, year, lung cancer incidence rate, race, sex, and concentration of the six major pollutants. Although it is possible to filter the data within the visualization software, these pre-process greatly reduces the size of the data file and makes it much easier to handle the data.

Data Analysis and Visualization
We used TIBCO Spotfire to analysis and visualize our data. We have utilized advanced information visualization methods such as Treemap, Heatmap and high dimensional scatter plot to explore the underlying trend in the dataset. We have also implemented coordinated windows and trellised view during this process.

Results
1. When it comes to lung cancer, the map is not flat

We plot treemap and heatmap of the average lung cancer incidence rate across United States over the period 2000 – 2009. Cancer incidence rate [5] is defined as the number of new cancers of a specific site/type occurring in a specified population during a year:

\[
\text{Incidence Rate} = \frac{\text{New Cancers}}{\text{Population}} \times 100,000
\]
2. Men are more susceptible to lung cancer, but it is getting better
Figure 2. Racial and gender differences of lung cancer. (A) A hierarchical treemap shows how lung cancer incidence differs by gender and races. (B-C). Change of lung cancer incidence rate over time for different races (B) and different gender (C).

We then explored the racial and gender differences of lung cancer. First we showed an overview of lung cancer rate with regards to race and gender using treemap (Fig. 2A). Block size is determined by population and color is related with cancer rate value: darker the color, higher the value.

From the graph it is interesting to see that races are split into two groups. One with higher cancer rate around 80 (American Indian/Alaska Native, Black, White), another group with lower rate around 40 (Hispanic, Asian/Pacific Islander). Their rates are slowly decreasing over time, but the difference remains (Fig. 2B).

We also noticed that male has a much higher cancer rate than female. This is presented in the treemap as well: male's block is always darker than female's regardless of race. Being already at a relatively lower level, the cancer rate of females is more stable over time. Conversely, males’ rate has drawn down dramatically over the past decade, making the difference across gender smaller. According to a report from National Cancer Institute [6], this is consistent with reductions in tobacco consumption among men (Fig. 2C).

3. Air quality matters!
Correlation between Lung Cancer Incidence Rate and six major air pollutant

Figure 3. Correlation between lung cancer incidence rate and six major air pollutants (Carbon Monoxide, Lead, PM2.5, PM10, Nitrogen Dioxide and Sulfur Dioxide) in the year of 2006. The data is displayed with high-dimensional scatter plot. Each data point represents a different state (color coded). The size of the point represents the population of that state, and the X and Y dimensions are cancer incidence rate and average concentration of a particular air pollutant in a user-defined period. The air pollution data has unit of micrograms/cubic meter, measured at 25°C.

Given six major pollutants, Carbon Monoixide, Lead, PM2.5, PM10, Nitrogen Dioxide and Sulfur Dioxide, we investigate their correlation with cancer rate. We draw scatter plot of each state's air pollutant versus cancer rate. According to the graph, several pollutants show pretty obvious positive correlation (Fig. 3), in particular, PM2.5(PM, or Particulate Matter, is made up of a number of components, including acids, organic chemicals, metals, soil or dust particles, and allergens [7]. PM2.5 describes particulate matter that is 2.5 micrometers in diameter and smaller.). A positive relationship is also identified with Nitrogen Dioxide and Lead, however, the relationship is obscured by many data entries with very low level of these two pollutants.

After we visually identified the correlation, we then performed statistical analysis on the correlation between lung cancer and air pollutant (Fig. 4). Consistent with the visual analysis, PM2.5 is highly correlated with cancer rate ($R=0.49$, $p<10^{-6}$), the other two pollutants (Lead and Nitrogen Dioxide) are also positively correlated with lung cancer incidence rate ($R=0.16$ and $0.08$), but did not reach the significant level ($p>0.1$). However, if we filter out the states with very low level of these two pollutant, the correlation will be much stronger. For example, the correlation
coefficient increased from 0.16 to 0.45 for Lead if we only consider the 15 states with highest level of lead pollution, which is almost as high as the PM2.5. This pose a great example of the necessity of visualization, since with only statistical analysis, we would likely to conclude Lead is not a risk factor for lung cancer.

**Figure 4.** Statistical analysis of the correlation between air pollution and cancer rate, using the “Data Relationship” tool of SpotFire. In this coordinated window, we can easily filter data by using the control panel on the right, the statistics and the scatter plot will updated accordingly.

**Critique on software**

[Positive]
It is amazing how easy and intuitive to work with Spotfire. The design of Spotfire closely follows many of the eight Golden Rules of Interface Design [8]. For example, the labels are highly consistent when we create multiple figures (e.g., the color code of states). It permits easy reversal of actions, and has shortcuts for many of its function. Spotfire also illustrate the simple, yet effective principle of “Overview first, zoom and filter, details on demand”. It contains many visualization functions for overview (e.g., treemap, heatmap, high-D scatter plot), as well as functions for detailed analysis (e.g., data relationship). The filters are also very helpful to zoom in on a particular part of the data.

[Negative]
1. Error Handling. We feel the error handling feature could be improved. When an error happens, we did not get much feedback from Spotfire, other than a message telling us which in-built function of Spotfire throws the error. On the
other hand, some quick diagnosis and reports on what is likely the problem will be much more helpful for the user.

2. Figure Export. We have encountered some difficulties to edit the exported figures (in pdf format) with Adobe Illustrator. In particular, the text labels/legends are fragmented sometimes, with each character treated as a separate text object, which makes it difficult to adjust the font size/style.

3. Statistical Analysis. Although information visualization is not about rigorous statistical analysis, it would still be helpful to extract some basic statistics from the data (e.g. p-value, goodness-of-fit). We feel the statistical analysis features of spotfire are relatively weak and limited. For example, the data relationship tool only provides linear regress, while it is also helpful some times to fit the data with other functions (e.g. quadratic).

4. Multiple Filters. Occasionally, we would like to apply different filters to generate different figures. For example, it will be helpful to show the trend over time average over all the other factors, as well as the data for a particular year of interest at the same time. We have to open two Spotfire instances and filter data differently to realize this. It would be helpful to allow multiple data filters in the same window.

5. The zoom-in feature is disabled by default and not convenient to use. It will be better if the user can use scroll wheel on mouse to direct zoom-in/out.

Conclusion

In this report, we have utilized the state-of-the-art information visualization techniques to explore the potential risk factors for lung cancer. Significant correlations between lung cancer incidence rate and air pollutant concentrations have been spotted first by visualization then by statistical analysis. Geographical characteristics of lung cancer incidence rate have also been presented with the aid of treemap and heatmap.
Reference


