HHS Public Access

Author manuscript

IEEE EMBS Int Conf Biomed Health Inform. Author manuscript; available in PMC 2017 May 04.

Published in final edited form as:

IEEE EMBS Int Conf Biomed Health Inform. 2016 February ; 2016: 565–568. doi:10.1109/BHI.2016.7455960.

Predicting Lung Cancer Incidence from Air Pollution Exposures Using Shapelet-based Time Series Analysis

Hong-Jun Yoon,

Oak Ridge National Laboratory, Oak Ridge, TN 37831 USA phone: 865-241-2626; fax: 865-574-6275

Songhua Xu, and

Information Systems Department, New Jersey Institute of Technology, Newark, NJ 07102 USA

Georgia Tourassi [Member, IEEE]

Oak Ridge National Laboratory, Oak Ridge, TN 37831 USA phone: 865-576-4829; fax: 865-574-6275

Abstract

In this paper we investigated whether the geographical variation of lung cancer incidence can be predicted through examining the spatiotemporal trend of particulate matter air pollution levels. Regional trends of air pollution levels were analyzed by a novel shapelet-based time series analysis technique. First, we identified U.S. counties with reportedly high and low lung cancer incidence between 2008 and 2012 via the State Cancer Profiles provided by the National Cancer Institute. Then, we collected particulate matter exposure levels ($PM_{2.5}$ and PM_{10}) of the counties for the previous decade (1998–2007) via the AirData dataset provided by the Environmental Protection Agency. Using shapelet-based time series pattern mining, regional environmental exposure profiles were examined to identify frequently occurring sequential exposure patterns. Finally, a binary classifier was designed to predict whether a U.S. region is expected to experience high lung cancer incidence based on the region's $PM_{2.5}$ and PM_{10} exposure the decade prior. The study confirmed the association between prolonged PM exposure and lung cancer risk. In addition, the study findings suggest that not only cumulative exposure levels but also the temporal variability of PM exposure influence lung cancer risk.

I. Introduction

There are several well documented lung cancer risk factors, including age, genetic profile, lifestyle choices, and environmental factors. A well known environmental risk factor for lung cancer is ambient air pollution. Typical sources contributing to ambient air pollution are combustion engines of vehicles, residential heating and industrial plants, local population,

This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a nonexclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (<http://energy.gov/downloads/doe-public-accessplan>).

seasonal changes, and local weather conditions. Observational studies have shown that an individual's lung cancer risk is influenced adversely by cumulated exposure to ambient air pollution [1]. However, estimation of the effects of environmental factors on cancer incidence is a difficult task because cancer is a long latency disease. To perform a case-control observational study of the effect of an environmental exposure factor on cancer risk, a researcher needs access to long-term records of environmental exposure for large number of subjects who developed clinically detectable cancer as well as exposure profiles of cancer-free subjects for careful comparison [2].

In this study we investigated whether monitoring the temporal environmental exposure profile of a geographic region can predict an expected rise in cancer incidence. Therefore, instead of studying the exposure profiles of individual subjects, we propose to study the exposure profiles of geographical regions. As proof-of-concept, we use lung cancer risk and its association to prolonged exposure to particulate matter (PM). The objective of the study was to investigate if there exists correlation between the lung cancer incidence rate of a geographical region and its history of PM exposure. Two openly available data sources were used to perform the study. The first dataset was AirData provided from the U.S. Environmental Protection Agency (EPA) [3]. EPA monitors outdoor air quality across the United States since 1958 for various pollutants. The data is stored along with geographical information of the monitoring site, and date of measurement and sample duration, which enable us to derive aggregate measures of the air pollutants. The second information source was from the National Cancer Institute, namely the State Cancer Profiles dataset [4], which provides surveillance statistics of various cancer types and profiles for U.S. states and counties. The NCI surveillance data include incidence, prevalence and mortality of cancers. This study focused on the cancer surveillance statistics for 2008–2012 to identify which U.S. counties suffered higher incidence of lung cancer.

The paper is organized as follows. Section II describes the air pollutants considered in this study and the time series analysis we applied based on shapelets [5] to identify temporal exposure patterns associated with significantly increased or decreased lung cancer incidence. The experimental design and analysis were done at the U.S. county level to identify geospatial exposure patterns that are expected to significantly impact future lung cancer incidence. Section III presents the study findings while Section IV provides discussion of the findings and future research directions.

II. Methods

Association between an individual's exposure to airborne pollutants and respiratory diseases has been well established [6]. In this paper we pursue a population-level investigation of the association of geolocal trends of air pollution to changes in the yearly incidence disease rate. Specifically we focus on lung cancer, a chronic disease well known to be associated with poor air quality.

A. Local Trends of Particulate Matter Levels

Particulate Matter (PM) is a complex mixture of suspended pollutants in a form of small particles and liquid droplets. EPA groups PM into two categories, mainly inhalable coarse

particles smaller than 10 micrometers in diameter and fine particles smaller than 2.5 micrometers in diameter. The former is named PM_{10} and the latter is known as $PM_{2.5}$. Primary particles are emitted directly from a source, while secondary particles form indirectly with chemicals such as sulfur dioxides and nitrogen oxides emitted from fuel combustion. Coarse particles are made up of primary particles, while fine particles are made up from secondary particles. Small particles can get deep into the lungs which may cause health issues. The International Agency for Research on Cancer (IARC) has classified PM as Group I carcinogen [7]. EPA categorizes Air Quality Index (AQI) values according to PM levels, which suggests $PM_{10} < 55 \mu\text{g}/\text{m}^3$ is good condition, $55\sim154 \mu\text{g}/\text{m}^3$ is moderate, $155\sim254 \mu\text{g}/\text{m}^3$ is unhealthy for sensitive groups, and $> 254 \mu\text{g}/\text{m}^3$ is an unhealthy condition. For $PM_{2.5}$, $0\sim12.0 \mu\text{g}/\text{m}^3$ is considered a good condition, $12.1\sim35.4 \mu\text{g}/\text{m}^3$ is moderate, $35.5\sim55.4 \mu\text{g}/\text{m}^3$ is unhealthy for sensitive groups, and $> 55.5 \mu\text{g}/\text{m}^3$ is considered unhealthy [8]. EPA has been monitoring PM exposure across the U.S since 1982. Specifically EPA provides data of hourly PM_{10} (total of 0~10 micrometers on standard temperature and pressure (STP)) exposure metric across 3,356 distinct sites in the United States and hourly $PM_{2.5}$ STP and local condition from 2,398 distinct sites. This data is available at [3].

In this study, we calculated monthly average PM exposure levels for all U.S. counties since 1982. The county-level estimation was based on inverse distance weighted averaging of the monthly estimates of all monitoring sites within 50 miles of radius from the geolocation center of the specific county. If no monitoring site within the radius was found, the exposure level for the particular timeframe was considered unspecified.

B. Lung Cancer Incidence

Relative lung and bronchus cancer risks of U.S. counties were determined by observing the lung cancer incidence rates found in the State Cancer Profiles provided by the National Cancer Institute [4]. For this study, we used the data reported in the table of annual incidence rates (cases per 100,000 people per year) for the period between 2008 and 2012. Rates were age-adjusted by the 2000 U.S. standard population. Counties with lung cancer incidence rates higher than the national average, 63.7, plus the 95% confidence interval were considered high risk. Similarly, counties with lung cancer incidence rates lower than the national average, 63.7, minus the 95% confidence interval were considered low risk. Counties with lung cancer incidence within the 95% confidence interval of the national average were considered average risk.

PM exposure levels of the high risk counties as well as low risk counties were then collected. Monthly exposure data from 1998 to 2007 comprised the exposure profile for each county (i.e., $10 \text{ years} \times 12 \text{ mos/yr} = 120$ average exposure measurements). If no data were available to infer exposure in the desired time frame, the county was excluded from the study. There were 276 counties for which we were unable to infer PM_{10} exposure and 194 counties for which we had no available data to infer $PM_{2.5}$ exposure. Figure 1 demonstrates an example of monthly average exposure of PM_{10} Total and $PM_{2.5}$ obtained as described above. Note that there exist few discrepancies where $PM_{2.5}$ levels exceed PM_{10} Totals, mainly due to the interpolation of missing data.

C. Time Series Analysis Using Shapelets

The objective of the study is to predict regional lung cancer risk at the U.S. county level by applying time series analysis of historical PM exposure levels. Overall, we performed the following step sequence. First, we performed frequent pattern mining of PM exposure time series to identify salient patterns associated with high lung cancer incidence. Then, we designed a classifier to predict lung cancer risk using these salient patterns.

There are several approaches to perform time series analysis, but for this study we adopted the shapelet-based time series mining. This is a comprehensive way to explore local shape similarity of time series data independent to phase, which provided self-interpretable characteristics to the time series data [5].

For a given set of time series T of length m , there exists many subsequences S of length l , where $l \leq m$, $l \geq MINLEN$, and $l \leq MAXLEN$. Note that any subsequence with unspecified exposure levels is not considered as a shapelet. We then collect maximally informative shapelets S_{opt} from S with respect to their level of support, which is the ratio of the time series that contains S_j to the total number of time series. Presence of the shapelet in the time series is established if the minimum distance between the given shapelet S_j and any subsequence of a time series T is less than $MINDIST$, where the distance measure is

$$dist(S_1, S_2) = \sqrt{\frac{1}{l} \sum_{k=1}^l (s_{1,k} - s_{2,k})^2}.$$

For a given training time series with truth label (T_b, y_b) , binary feature vector $(\bar{x}_b, y_b)$ for time series classification based on frequent shapelets was composed where $\bar{x} = (x_1, x_2, \dots, x_k)$ and k is the size of S_{opt} . $x_r = 1$ if $S_r \in S_{opt}$ and $S_r \subseteq T_b$ and $x_r = 0$ otherwise for $r \in [1, k]$. The training dataset is then fed to a Random Forest classifier implemented using the Weka software package [9].

III. Experimental Results

According to the State Cancer Profiles provided by the National Cancer Institute, there were 1,008 U.S. counties with significantly higher annual lung cancer incidence rates from year 2008 to 2012, and 348 counties with significantly lower annual lung cancer incidence rates for the same timeframe. Our experiment was designed to examine whether the shapelet-based features from the local air pollution exposure levels of the previous 10 years (1997–2007) could discriminate between high- and low-risk counties in terms of their lung cancer incidence rate reported by NCI for the period 2008–2012.

Some key statistics of the monthly average PM exposures are listed in Table I, which exhibits few differences of average exposures of PMs between the U.S. counties of the high and low risk groups. Classification of high and low risk counties based on average PM₁₀ exposures results in ROC area of 0.569, suggesting that average PM₁₀ exposure levels experienced in the U.S. are controlled well enough and lack predictive power. In contrast, classification accuracy based on average PM_{2.5} exposures results in substantially higher

ROC area of 0.729, suggesting that average $PM_{2.5}$ exposure levels experienced in the U.S. require further improvement.

From Figure 1, we note that within a geographical region there exist substantial monthly fluctuations which may be due to noise with the measurement/data collection process. To mitigate the effect of possible measurement noise while preserving seasonal change of air quality, we performed experiments with average exposures of 3-month and 6-month periods along with monthly average exposures. Optimal shapelets of the PM exposures were determined with *MINLEN* and *MAXLEN* to capture 1 to 5 years of PM exposure trends, i.e., for monthly average exposures we set *MINLEN* = 12 and *MAXLEN* = 60, and for 6-month average exposures we set *MINLEN* = 2 and *MAXLEN* = 10. We selected $k = 50$ shapelets based on their level of support. Oversampling was applied to the data of low risk counties to mitigate class imbalance. Then, a Random Forest classifier was developed using the 50 shapelets to discriminate between high and low lung cancer risk counties. 10-fold cross validation results are listed in Table II.

The results in Table II demonstrate that both PM_{10} and $PM_{2.5}$ -based shapelet features possess discrimination power. These results confirm the association between higher exposure to PM and higher risk of lung cancer. Note that the ROC areas of the shapelet-based classifiers are substantially higher than the ROC areas of classification based on average PM exposures. This finding demonstrates that the shapelet is an effective feature representation scheme.

Classification based upon 3-month and 6-month average exposures of both PM's provide better predictive power than using monthly average, with the 3-month average performing better than the other two methods for PM_{10} and comparably to using 6-month average for $PM_{2.5}$. These results suggest suppressing data collection noise while maintaining seasonal characteristics of PM exposures is an effective approach.

Table III lists the 12 shapelets from the dataset of 3-month average PM exposures that represent strong association to high and low lung cancer risk. The first 6 shapelets listed are from analyzing the temporal PM_{10} exposure profiles of the U.S. counties, while the remaining 6 are from analyzing the temporal $PM_{2.5}$ exposures. Shapelets of solid lines are the ones associated with high lung cancer risk, and those of dotted lines are associated with low lung cancer risk. Average exposures of PM and associated odds ratio to lung cancer risk are listed along.

A notable difference between the shapelets associated with high lung cancer risk and those associated with low lung cancer risk was observed. Namely, the shapelets associated with high lung cancer risk demonstrate higher fluctuation, while the low cancer risk shapelets demonstrate a relatively steady exposure pattern. This finding is consistent for both PM_{10} and $PM_{2.5}$ exposures suggesting that not only the amount of exposure but also the temporal variation impact future cancer risk.

IV. Discussion

In this study, we proposed a method to predict whether a geographical region is expected to experience significantly elevated cancer incidence based on its environmental exposure profile during the prior decade. For proof of concept we used lung cancer and exposure to airborne particulate matter which is known to be carcinogenic. Our method leverages shapelet-based time series analysis of U.S. counties' local exposure to PM₁₀ and PM_{2.5}. Overall, our study findings were the following.

- Experimental results confirmed that the prolonged high exposure of PM₁₀ and PM_{2.5} adversely influences lung cancer risk.
- Analysis of individual shapelets suggests an association between high lung cancer risk and highly fluctuating PM exposures, and association of low lung cancer risk with steady PM exposures.
- Analysis of individual shapelets further suggests that high risk PM_{2.5} shapelets are associated with higher PM_{2.5} exposure levels than the low risk ones. In contrast, the high risk PM₁₀ shapelets exhibit marginal difference of exposure level from their low-risk counterparts.

Obviously, PM is not the sole risk factor for lung cancer. To achieve better understanding and prediction of lung cancer or other types of cancer risk, it is strongly suggested to apply time series of multiple channels of various environmental, lifestyle, socio-economic and other factors that may be associated with cancer risk. The proposed method can be extended to accommodate additional factors.

In conclusion, the main advantage of the overall approach is its ability to discover carcinogenic factors leveraging publicly available aggregate data collected by cancer surveillance programs and other national programs. This is a cost-effective way for knowledge discovery in environmental cancer epidemiology to prioritize and guide the design of more precise but also more time-consuming patient-level observational studies.

Acknowledgments

This study was supported by Grant No. 1R01-CA170508-03 from the National Cancer Institute.

References

1. Stewart, BW., Kleihues, P., editors. World Cancer Report. Vol. 57. Lyon: IARC Press; 2003.
2. Vineis P, et al. Air pollution and risk of lung cancer in a prospective study in Europe. International journal of cancer. 2006; 119(1):169–174. [PubMed: 16463382]
3. United States Environmental Protection Agency. AirData. 2015 Nov 24. [Online]. Available: <http://www3.epa.gov/airdata/>
4. National Cancer Institute. State Cancer Profiles. 2015 Nov 24. [Online]. Available: <http://statecancerprofiles.cancer.gov/>
5. Ye, L., Keogh, E. Time series shapelets: a new primitive for data mining. Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining; ACM; 2009. p. 947-956.
6. Brunekreef B, Holgate ST. Air pollution and health. The lancet. 2002; 360(9341):1233–1242.

7. Loomis D, Grosse Y, Lauby-Secretan B, El Ghissassi F, Bouvard V, Benbrahim-Tallaa L, Guha N, Baan R, Mattock H, Straif K. The carcinogenicity of outdoor air pollution. *The lancet oncology*. 2013; 14(13):1262–1263. [PubMed: 25035875]
8. United States Environmental Protection Agency. Technical Assistance Document for the Reporting of Daily Air Quality – the Air Quality Index (AQI). 2013. EPA-454/B-13-001
9. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The Weka data mining software: an update. *ACM SIGKDD explorations newsletter*. 2009; 11(1):10–18.
10. United States Environmental Protection Agency. National Ambient Air Quality Standards (NAAQS). 2015 Dec 1. [Online]. Available: <http://www3.epa.gov/ttn/naaqs/criteria.html>

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 1.
Monthly average exposure of PM₁₀ Total 0~10 micrometers and PM_{2.5} in Milwaukee County, Wisconsin from year 1998 to 2007.

TABLE I

Statistics of Average Exposures of PM₁₀ and PM_{2.5} (Micrograms per Cubic Meter) from 1998 to 2007 of U.S. Counties of High and Low Risks of Lung Cancer according to the Annual Lung Cancer Incidence from 2008 to 2012

	Risk	Min.	Max.	Mean	Std.
PM ₁₀	High	9.66	104.65	22.52	4.68
	Low	4.89	51.38	23.24	6.15
PM _{2.5}	High	5.08	17.30	11.76	2.00
	Low	2.45	18.51	9.37	3.35

Classification Performance of Lung Cancer Risks of U.S. Counties from Monthly Average Exposures of PM_{10} and $PM_{2.5}$ based on Shapelet Analysis.
Exposure Monitored from Year 1998 to 2007

TABLE II

Period (month)	TP Rate	FP Rate	Precision	Recall	AUC
<i>PM₁₀</i>					
1	0.774	0.518	0.752	0.774	0.747
3	0.783	0.496	0.764	0.783	0.766
6	0.758	0.550	0.729	0.758	0.735
<i>PM_{2.5}</i>					
1	0.827	0.325	0.821	0.827	0.824
3	0.829	0.349	0.821	0.829	0.833
6	0.814	0.363	0.805	0.814	0.835

TABLE III

Individual Shapelets of PM_{10} and $PM_{2.5}$ 3-month Average Exposure Levels associated with High Lung Cancer Risk (in Solid Lines) and Low Lung Cancer Risk (in Dotted Lines) with Average Exposure values and Odds Ratios to Lung Cancer Risk. Row S for Shapelets, M for Averages and Standard Deviations, and O for Odds Ratios and Confidence Intervals

PM_{10}			
S			
M	21.24 (4.34)	22.21 (5.55)	21.15 (6.01)
O	6.64 (3.38, 11.36)	5.64 (3.52, 9.05)	4.44 (2.91, 6.78)
S			
M	23.01 (2.85)	20.44 (5.55)	18.24 (3.60)
O	0.30 (0.20, 0.46)	0.35 (0.25, 0.51)	0.36 (0.25, 0.51)
$PM_{2.5}$			
S			
M	13.67 (3.56)	13.86 (2.87)	14.30 (2.75)
O	11.58 (8.72, 15.39)	11.57 (8.72, 15.34)	11.15 (8.39, 14.80)
S			
M	4.66 (1.34)	4.76 (1.42)	6.38 (2.69)
O	0.07 (0.04, 0.10)	0.07 (0.05, 0.11)	0.08 (0.05, 0.12)