International Journal of Biomedical Science and Engineering
2017; 5(3): 18-23
http://www.sciencepublishinggroup.com/j/ijbse
doi: 10.11648/j.ijbse.20170503.11
ISSN: 2376-7227 (Print); ISSN: 2376-7235 (Online)

Identifying Air Pollution Risk Factors for Respiratory Disease Using Quantitative Computational Method

Songjing Chen, Sizhu Wu, Qing Qian, Jiao Li*

Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing, China

Email address:
li.jiao@imicams.ac.cn (Jiao Li)
*Corresponding author

To cite this article:
Songjing Chen, Sizhu Wu, Qing Qian, Jiao Li. Identifying Air Pollution Risk Factors for Respiratory Disease Using Quantitative Computational Method. *International Journal of Biomedical Science and Engineering.* Vol. 5, No. 3, 2017, pp. 18-23.
doi: 10.11648/j.ijbse.20170503.11

Received: April 1, 2017; **Accepted**: April 24, 2017; **Published**: May 5, 2017

Abstract: In order to identify air pollution risk factors for respiratory disease patients, a quantitative computational method to identify high risk factors for respiratory patients was conducted in this study. The C4.5 classification algorithm was used in the computational method. SMOTE algorithm was applied to solve the imbalance data problem. Risk factor effect degree was calculated according to C4.5 classification model. Age was the top risk factor in nine subgroups, except ≤11. The age ≤49 youth were easier affected by NO_2 and SO_2 than >49. ≤49 were obviously more than >49. ≤49 were more easier suffer from acute upper respiratory infections, >49 were more easier suffer from influenza, pneumonia and chronic lower respiratory disease. The air pollution risk factors of respiratory disease were identified quantitatively. This quantitative computational method could be applied to predict other disease occurrence.

Keywords: Respiratory Disease, Air Pollutant, C4.5 Classification Algorithm, Data Mining

1. Introduction

Air pollution causing respiratory disease is becoming serious nowadays. The high level of air pollutants may increase respiratory admissions. This association suggests a possible link between respiratory disease occurrence and exposure to air pollution, in particular, to PM10, SO_2 and NO_2. Many studies were conducted in Europe and North America [1] in this area. Air Pollution and Health: A European Approach (APHEA) [2] shows that in the warm season, an increase in the 1-hour ozone concentration by 10μg/m3 is associated with a 0.33% increase in the total daily number of deaths. The National Morbidity, Mortality, and Air Pollution Study (NMMAPS) conducts in the 90 largest American cities and analyzes a 10μg/m3 increase in the previous day's PM10 is associated with an approximate 0.2% increase in daily mortality [4]. The Public Health and Air Pollution in Asia (PAPA) program [5], conducts a time-series study in Shanghai, China to investigate the relationship between outdoor air pollution and daily mortality from 2001 to 2004 using 4 years of daily data. Some targeted factor is discussed in these studies, and the emphasis of this research is to compute effect degree of different air pollution risk factors quantitatively.

The present study is to extract risk factors of respiratory disease in different stratified groups. According to the population characteristics, targeted risk factors are extracted and quantitative analysis is conducted. This quantitative computational method could be applied to predict other disease occurrence.

2. Materials

In this research, 135,008 daily emergency admissions were used. These data were from three large-scale comprehensive hospitals in Beijing, China, between January 1, 2009 and December 31, 2011. These three hospitals were almost the top level hospitals in China. Admissions were outpatients in respiratory emergency room and were described in Figure 1.

The environmental data were daily measured in the Beijing Environmental Monitoring Centre from January 1, 2009 to December 31, 2011. Nitrogen dioxide (NO_2), sculpture dioxide (SO_2) and particulate mass less than 10 microns in

aerodynamic diameter (PM10) were obtained. The meteorological data such as temperature, relative humidity, wind speed, atmospheric pressure, sunshine time and rainfall precipitation were also monitored. These environmental risk were measured in Beijing Environmental Monitoring Centre.

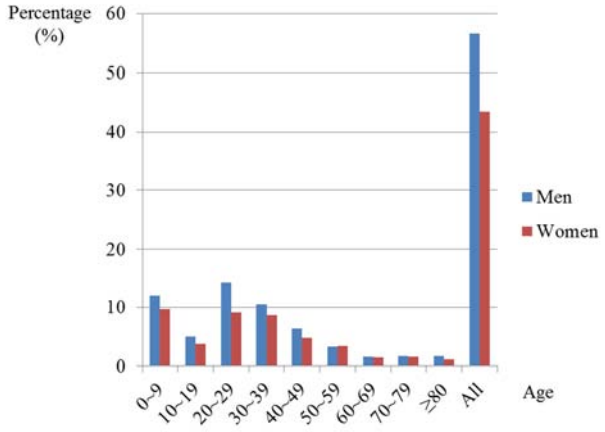


Figure 1. *The histograms of sex and age distribution.*

3. Methods

The key risk factors were extracted from numerous factors in different groups. Based on the emergency patient data and the environmental data, a risk factors identification method was adopted in this study. A computational method was used to realize the extraction model, and machine learning and statistical analysis were applied to construct the model. The emergency patient data were classified into two classes in the data pre-processing stage. In order to solve the imbalanced data problem, the SMOTE (Synthetic Minority Over-sampling Technique) algorithm [6, 7, 8] was adopted. And then the entire population was divided into eight subgroups by sex and age. There were nine groups in total, including the entire group and eight subgroups. Then the extraction model was generated using the C4.5 classification algorithm. And risk factors were identified according to decision trees.

3.1. Data Processing

Data cleaning and data classifying were included in the data pre-processing stage.

3.1.1. Data Cleaning

There were vacancies, noise, and unexpected semantic information and so on in the clinical data, these might cause error on research results. So, at first, we filled vacant values, identified isolated points, abated the noise and corrected inconsistencies in the data. Then duplicate elements and excessive vacancy factors were removed through data protocol. At last, 30 dimensions 135,008 records were obtained, including disease conditions, patient's physical condition and environmental risk factors. Disease condition was diagnose result. Patient's physical condition included sex and age. Environmental risk factors were nitrogen dioxide (NO_2), sculpture dioxide (SO_2), particulate mass less than 10 microns in aerodynamic diameter (PM10), 24-hour average temperature, daily maximum temperature, daily minimum temperature, relative humidity, wind speed, daily maximum wind speed, daily minimum wind seed, air pressure, daily maximum air pressure, daily minimum air pressure, sunshine time, rainfall precipitation and season and so on.

3.1.2. Data Classification

According to clinical diagnosis, the International Classification of Diseases, Tenth Revision (ICD-10) [9] were used to classify these data. There were acute upper respiratory infections (J00-J06), influenza and pneumonia (J09-J18), chronic lower respiratory diseases (J40-J47) and others. In this research, acute upper respiratory infections (J00-J06) were defined as Diagnosis 1, occupying 88% in the entire outpatients. Influenza and pneumonia (J09-J18) (occupying 6%), chronic lower respiratory diseases (J40-J47) (occupying 46%) and others were defined as Diagnosis 2, occupying 12% in total. The ratio of Diagnosis 1 to Diagnosis 2 was about 7. So Diagnosis 1 was too more than Diagnosis 2, and the data existed imbalance problem.

3.2. Imbalanced Data Solution

SMOTE (Synthetic Minority Over-sampling Technique) algorithm was adopted to solve the imbalanced data problem. SMOTE was the oversampling method by increasing the samples of the minority class to improve the classifier performance.

The classifiers' performance without using SMOTE was shown in Table 1. And after using SMOTE algorithm, the classifiers' performance was list in Table 2. So comparing these two tables, the improvement of classifiers' performance could be seen easily.

Table 1. *Classifiers performance before using SMOTE.*

Classifiers	Precision (%)	ROC
C4.5	78.05%	0.68
Random Forest	82.6%	0.73
Logistic Regression	83.2%	0.64
Bayes Net	79.81%	0.75
RBF Network	80.56%	0.77

Table 2. *Classifiers performance after using SMOTE.*

Classifiers	Precision (%)	ROC
C4.5	89.8%	0.86
Random Forest	85.6%	0.83
Logistic Regression	86.1%	0.84
Bayes Net	80.56%	0.76
RBF Network	78.55%	0.73

3.3. Groups Division

To conduct targeted quantitative analysis in different crowds, the entire population was divided into eight subgroups according to the sex and the age. The decision tree was shown in Figure 2, age and sex were near the root node of tree. And the age was divided by 49, then by 11 and 68, and so on. We chose the obvious characteristics to divide the entire group into eight subgroups. The eight subgroups contented age >49, age ≤49,

men >49, women >49, men ≤49, women ≤49, men ≤11 and women ≤11. There were nine groups in total.

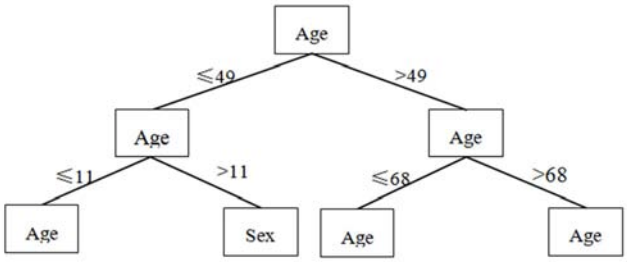


Figure 2. *The upper part of the entire group decision tree.*

3.4. Risk Factors Identification

The risk factors in different subgroups were extracted respectively. It was consisted of three main steps: C4.5 model training, the effect degree (given in Equation (1)) calculation of risk factors, effect degree analysis.

3.4.1. C4.5 Model Training

We used C4.5 classifier algorithm to train classification model, and adopted the 10-fold cross validation to verify the trained model. Set the entire group as an example, at first, the training data were used to train the C4.5 model. And the 10-fold cross validation was used. And then the performance of C4.5 model was analyzed. The decision tree was obtained. Weka 3.5.8 program was used to train the classification model. The training parameters were Confidence Factor = 0.25, the minimum number of instance per leaf = 2.

Therefore, the eight subgroups were separately trained to generate classification models in the same way. The performance of these C4.5 models in different groups was given in Table 3.

Table 3. *The performance of C4.5 models in different groups.*

Performance / Models	Precision (%)	ROC
All	87.01	0.876
Age>49	85.99	0.757
Age≤49	79.82	0.858
Age>49 Women	84.44	0.773
Age>49 Men	86.14	0.731
Age≤49 Women	80.98	0.87
Age≤49 Men	78.44	0.847
Age≤11 Women	92.32	0.882
Age≤11 Men	90.63	0.854

3.4.2. Factors Effect Degree Calculation

According to the decision tree, the top four layers were used to calculate the effect degree. The calculation formula of the risk factor effect degree was shown in Equation (1).

$$\delta(L,n)=\sum_{L=1}^{j}\sum_{n=1}^{i} n\times\left(\frac{1}{2}\right)^{L-1} \quad (1)$$

L: the Lth layer in the decision tree

n: the occurrence number in the Lth layer

Decision trees which were obtained from the entire group and eight subgroups were used to calculate the effect degree separately.

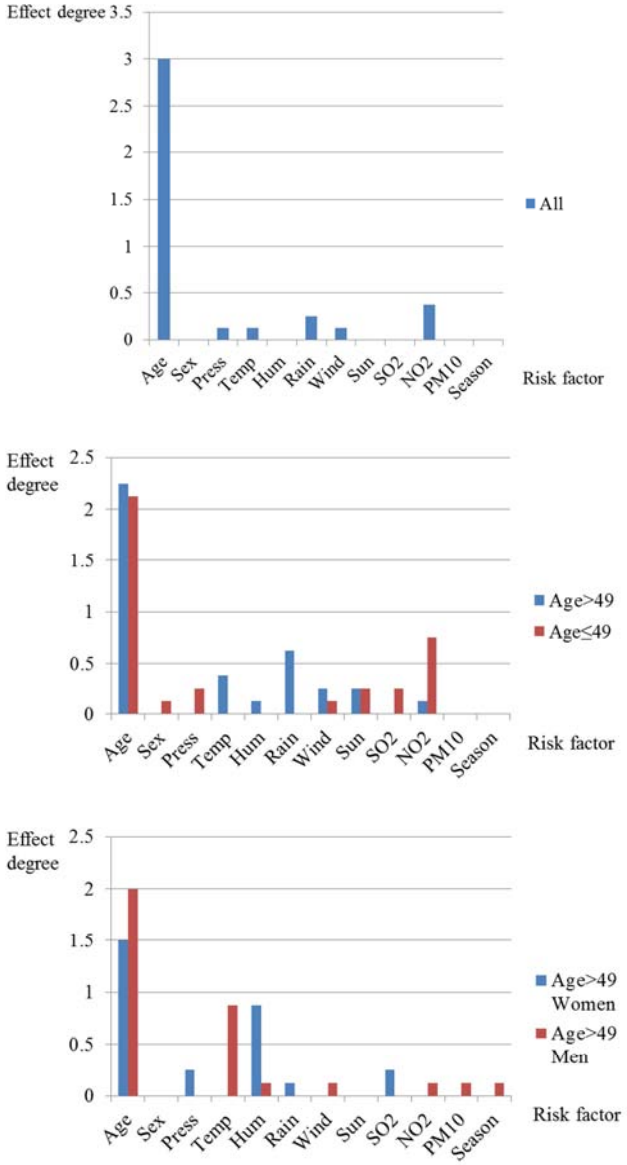
3.4.3. Effect Degree Analysis

These effect degrees of different groups were reorganized, and a series of statistical analysis was conducted. And the analysis results were given in Results section.

4. Results

4.1. Effect Degrees of Risk Factors

The effect degree of risk factors in the entire group and the eight subgroups was obtained respectively. And the histograms of these groups were shown in Figure 3.



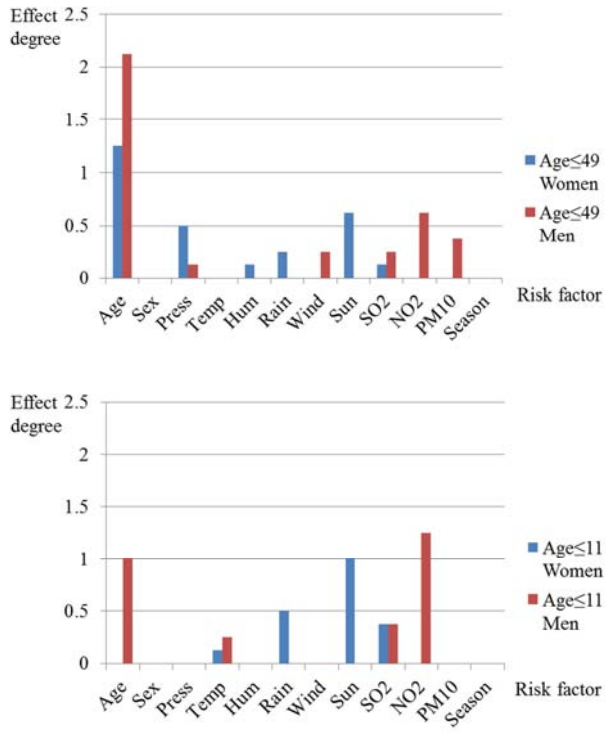


Figure 3. *The histograms of effect degrees in these nine groups.*

Through effect degree analysis, the following results could be acquired:

(1) The age was the top risk factor which caused respiratory disease. In these nine groups, there were seven groups the age effect degree lying in the top one, except two subgroups in ≤11.

(2) The younger group (age ≤49) was easier affected by air pollutants, such as NO_2 and SO_2, than the older group (age >49). The effect degree of NO_2 in ≤49 group was five times larger than >49 group. But rainfall, temperature and wind velocity had larger effect degree in >49 group than ≤49 group. Rainfall and temperature had little effect on ≤49 younger group. In >49 group, the effect degree of wind velocity was two times as large as ≤49 group.

(3) These NO_2, PM10 and SO_2 environmental risk factors had more effect on ≤49 men than ≤49 women. The effect of NO2 and PM10 on ≤49 men group was obviously larger than their effect on ≤49 women. And the effect of NO_2 and PM10 was little on ≤49 women. The effect of SO_2 on ≤49 men was two times as large as ≤49 women. The sunshine time had more effect on ≤49 women, which had almost no effect on ≤49 men. The effect of atmospheric pressure on ≤49 women was three times larger than the effect on ≤49 men.

(4) SO_2 had much more effect on >49 women than >49 men. And >49 men were easier affected by NO_2 and PM10 than >49 women.

(5) The sunshine time had much more effect on ≤11 girls than ≤11 boys. But ≤11 boys were easier affected by NO_2 than ≤11 girls.

4.2. Admission Numbers Analysis

Summary statistics of emergency room admissions in the two diagnosis classes were recorded in the nine groups respectively. The statistical results were shown in Table 4, and the histogram was given in Figure 4.

Table 4a. *The number of two classes patients in these nine groups.*

Classes	All	>49	≤49	>49Women	>49Men
Diagnosis 1	111798	11155	100643	5600	5555
Diagnosis 2	92482	39812	52670	17721	22091

Table 4b. *The number of two classes patients in these nine groups.*

Classes	≤49 Women	≤49 Men	≤11 Women	≤11 Men
Diagnosis 1	43350	57293	13769	17060
Diagnosis 2	21015	31655	1472	2334

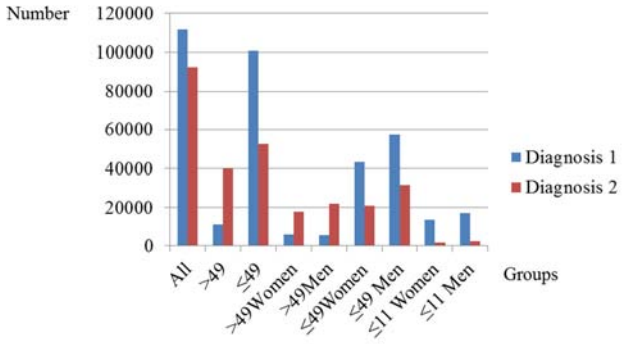


Figure 4. *The histogram of the patient quantities statistics.*

According to statistics results in the two classes, we could conclude the following items:

(1) The patient number in ≤49 group was obviously larger than >49 older group, especially in Diagnosis 1 (acute upper respiratory infections). In the Diagnosis 1, the ≤49 youth who suffered from acute upper respiratory infections was 8.02 times as many as the >49 group.

(2) In the ≤49 group, there were much more people suffering from acute upper respiratory infections than influenza, pneumonia and chronic lower respiratory diseases. But much more people in >49 suffer from influenza, pneumonia and chronic lower respiratory diseases than acute upper respiratory infections. The patients in ≤49 group who suffer from acute upper respiratory infections (Diagnosis 1) was 1.91 times as many as the patients who suffer from influenza, pneumonia and chronic lower respiratory diseases (Diagnosis 2). And in the >49 older group, the patients who suffer from influenza, pneumonia and chronic lower respiratory diseases (Diagnosis 2) were 3.57 times as many as the patients who suffer from acute upper respiratory infections (Diagnosis 1).

(3) In ≤11 groups, not only girls but boys, the most patients suffered from acute upper respiratory.

4.3. Evaluation

In order to evaluate this research method, meta-analysis was conducted for comparison. The meta-analysis was used commonly in this research area [10, 11, 12]. Classification using regression methods were done through the Weka 3.5.8 program. Set the entire group as an example, the top three

risk factors were age, NO_2 and rainfall obtained by the evaluation model. This order was the same as the C4.5 model, which was given in Figure 3. But the model precision was lower 12.70% than C4.5 model as shown in Table 5 Comparing with C4.5 model, shown in Table 3. And the ROC was also lower than C4.5 model. The other models were obtained, and the model performance was in Table 5.

Table 5. *The performance in different groups using meta-analysis.*

Performance / Models	Precision (%)	ROC
All	74.31	0.832
Age>49	81.08	0.752
Age≤49	72.02	0.785
Age>49 Women	79.43	0.757
Age>49 Men	82.77	0.744
Age≤49 Women	73.89	0.808
Age≤49 Men	71.16	0.778
Age≤11 Women	85.80	0.808
Age≤11 Men	86.94	0.791

5. Discussion

In this research, the environmental risk factor of respiratory disease in different groups was extracted respectively and measured quantitatively. This computational method was realized by C4.5 decision tree with high accuracy, which different from the previous studies. Meta-analysis [13], time-series [14, 15] and case-crossover [16, 17] were used more in these related works. The APHENA (Air Pollution and Health: A Combined European and North American Approach) study applied meta-regression approaches [2] and multi-cities time-series [18] to research the effect of air pollution on population health. Ling Tong [19] investigated association between air pollutants and cardiovascular morbidity using time-series analysis. Valerie B Haley [20] used time-stratified case-crossover to estimate differences in the short-term impacts of PM2.5 on cardiovascular disease hospital admissions in New York State.

The following conclusions could be summarized. 1) Age was the top risk factor in both the entire group and eight subgroups, except ≤11 group. 2) The≤49 youth were easier affected by NO_2 and SO_2 than >49 group. Especially, ≤49 men were much more sensitive to NO_2, PM10 and SO_2. 3) In the emergency room, patients who were ≤49 were obviously more than >49. And ≤49 were more easier suffer from acute upper respiratory infections, >49 were more easier suffer from influenza, pneumonia and chronic lower respiratory disease.

A number of researches focused on two or three environmental factors and analyzed their impact [21, 22, 23, 24]. The limited air pollutants were considered in this research, however in practice, once air pollution risk factors were controlled, the other pollutants would have been decreased as well. This made the risk factors more flexible in practice. Emergency patients associated with short-term expose to air pollution, and did not consider the adverse effects of chronic exposure to pollutants. We did not take into account long-term studies of air pollution in the current quantitative analysis. Therefore, additive temporal effects of air pollution were not quantified in this research. The confounding factors of emergency outpatients, such as smoke and chronic disease, were not collected in respiratory emergency room. And it was a shortage of this research.

The risk factor was extracted quantitatively in different groups. But we did not know the sensitive intervals of these risk factors, in which the risk factor decreased a small interval would cause more decreasing respiratory incidence. Therefore, in future work, this research can be conducted.

6. Conclusion

The air pollution risk factors of respiratory disease were identified quantitatively. This quantitative computational method could be applied to predict other disease occurrence. This computational method could be applied to predict the relationship between air pollution and respiratory occurrence using real-time data. The results could help clinicians understand the association between respiratory disease occurrences and expose to air pollution. The idea of population refinement could prompt that special population might suffer from corresponding disease in certain environmental condition. That might help clinicians take appropriate measures of patients visit in various environmental statuses.

Acknowledgements

The research is supported by the Fundamental Research Funds for the Central Universities (Grant No. 2016ZX330020), the National Population and Health Scientific Data Sharing Program of China, the Knowledge Centre for Engineering Sciences and Technology (Medical Centre), and the Key Laboratory of Knowledge Technology for Medical Integrative Publishing.

References

[1] Klea K, Jonathan M. Air pollution and health: A European and North American Approach (APHENA). Health Effects Institute Research Report, 142, 5-90(2009).

[2] Alexandros G, Bertil F, Klea K, et al. Acute effects of ozone on mortality from the "Air Pollution and Health: A European Approach" Project. American Journal of Respiratory and Critical Care Medicine. 170(10), 1080-1087 (2004).

[3] Evangelia S, Roger P, Tim R, et al. Acute effects of ambient particulate matter on mortality in Europe and North America: Results from the APHENA Study. Environmental Health Perspectives. 116(11), 1480-1486 (2008).

[4] Dominici F, McDermott A, Daniels M, et al. Revised analyses of the national morbidity, mortality, and air pollution study: Mortality among residents of 90 cities. Journal of Toxicology and Environmental Health. 68(13), 1071-1092 (2005).

[5] Kan H, Chen B, Zhao N, et al. Part 1. A time-series study of ambient air pollution and daily mortality in Shanghai, China. Health Effects Institute Research Report. 154, (2010)17-78.

[6] Nitesh V, Kevin W, Lawrence O, et al. SMOTE: Synthetic Minority Over-sampling technique. Journal of Artificial Intelligence Research. 12(6), 321-357 (2002).

[7] Nakamura M, Kajiwara Y, Otsuka A, et al. LVQ-SMOTE - Learning vector quantization based Synthetic Minority Over-sampling Technique for biomedical data. Biodata Mining. 6(16), 1-10 (2013).

[8] Dai, HL. Class imbalance learning via a fuzzy total margin based support vector machine. Applied Soft Computing. 31(1), 172-184(2015).

[9] ICD-10 Version:2010 http://apps.who.int/classifications/icd10/browse/2010/en.

[10] Anoop S, Jeremy P, Harish N, et al. Global association of air pollution and heart failure: a systematic review and meta-analysis. Lancet. 382(9), 1039-1048(2013).

[11] Mustafic H, Jabre P, Caussin C, et al. Main air pollutants and myocardial infarction: a systematic review and meta-analysis. The Journal of the American Medical Association. 307(7), 713–721 (2012).

[12] Duval S, Tweedie R. Trim and fill: a simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. Biometrics. 56(2), 455–463(2000).

[13] Majid E, Stephen V, Anthony R, et al. Estimates of global and regional potential health gains from reducing multiple major risk factors. Lancet. 362(7), 271-280(2003).

[14] Nicos M, Panayiotis Y, Savvas K, et al. A 10-year time-series analysis of respiratory and cardiovascular morbidity in Nicosia, Cyprus: the effect of short-term changes in air pollution and dust storms. Environmental Health. 39(7), 1-16(2008).

[15] Shan Zheng, Minzhen Wang, Shigong Wang, et al. Short-Term effects of gaseous pollutants and particulate matter on daily hospital admissions for Cardio-Cerebrovascular disease in Lanzhou: Evidence from a heavily polluted city in China. International Journal of Environmental Research and Public Health. 10(2), 462-477(2013).

[16] Monica C, Ennio C, Massimo S, et al. Short-Term effects of Nitrogen Dioxide on mortality and susceptibility factors in 10 Italian cities: The EpiAir Study. Environmental Health Perspectives. 119(11), 1233-1238(2011).

[17] Colais P, Faustini A, Stafoggia M, et al. Particulate air pollution and hospital admissions for cardiac diseases in potentially sensitive subgroups. Epidemiology. 23(3), 473–481(2012).

[18] Peng RD, Samoli E, Pham L, et al. Acute effects of ambient ozone on mortality in Europe and North America: results from the APHENA study. Air Qual Atmos Health. 6(2), 445-453(2013).

[19] Ling Tong, Kai Li, Qixing Zhou. Promoted relationship of cardiovascular morbidity with air pollutants in a typical Chinese urban area. Plos One. 9(9), 1-7(2014).

[20] Valerie B Haley, Thomas O, Henry D. Surveillance of the short-term impact of fine particle air pollution on cardiovascular disease hospitalizations in New York State. Environmental Health. 42(8), 1-10(2009).

[21] Richard T, Jeffery R. Effects of particulate and gaseous air pollution on cardiorespiratory hospitalizations. Archives of environmental health. 54(2), 130-139(1999).

[22] Leah J, Scott L. Are the acute effects of particulate matter on mortality in the national morbidity, mortality, and air pollution study the result of inadequate control for weather and season? A sensitivity analysis using flexible distributed lag models. American Journal of Epidemiology. 162(1), 80-88(2005).

[23] Douglas W, David Q, Patrick G, et al. Effect of air pollution control on mortality and hospital admissions in Ireland. Health Effects Institute Research Report. 176, 3-109(2013).

[24] Minzhen Wang, Shan Zheng, Shilin He, et al. The association between diurnal temperature range and emergency room admissions for cardiovascular, respiratory, digestive and genitourinary disease among the elderly: A time series study. Science of the Total Environment. 456(7), 370-375(2013).