Machine learning prediction on number of patient due to conjunctivitis based on air pollutants: A preliminary study

Juan Chen
Eye hospital ,Nanjing medical University
Yong-ran Cheng
Hangzhou Medical College
Zhan-hui Feng
Affiliated Hospital of Guizhou Medical University
Meng-Yun Zhou
Shinshu University School of Medicine
Nan Wang
First People's Hospital of Suzhou
Lan Ye
Guizhou Medical University
Mingwei Wang (wmw990556@163.com)
Hangzhou Normal University Affiliated Hospital https://orcid.org/0000-0001-9060-5107
Jin Yao
Eye Hospital,Nanjing Medical University

Research

Keywords: machine learning, patient for conjunctivitis, air pollutant

DOI: https://doi.org/10.21203/rs.3.rs-52822/v1

Abstract

Background: Accurate prediction of the number of patients with conjunctivitis plays an important role in providing adequate treatment at the hospital, but such accurate predictive model currently does not exist. The current study sought to use machine learning (ML) prediction based on past patient for conjunctivitis and several air pollutants. The optimal machine learning prediction model was selected to predict conjunctivitis-related number patients.

Methods: The average daily air pollutants concentrations (CO, O_3, NO_2, SO_2, PM10, PM2.5) and weather data (highest and lowest temperature) were collected. Data were randomly divided into training dataset and test dataset, and normalized mean square error (NMSE) was calculated by 10 fold cross validation, comparing between the ability of seven ML methods to predict the number of patient due to conjunctivitis (Lasso penalized liner model, Decision tree, Boosting regression, Bagging regression, Random forest, Support vector, and Neural network). According to the accuracy of impact prediction, the important air and weather factors that affect conjunctivitis were identified.

Results: A total of 84977 cases to treat conjunctivitis were obtained from the ophthalmology center of the Affiliated Hospital of Hangzhou Normal University. For all patients together, the NMSE of the different methods were as follows: Lasso penalized liner regression: 0.755, Decision tree: 0.710, Boosting regression: 0.616, Bagging regression: 0.615, Random forest: 0.392, Support vectors: 0.688, and Neural network: 0.476. Further analyses, stratified by gender and age at diagnosis, supported Random forest as being superior to others ML methods. The main factors affecting conjunctivitis were: O_3, NO_2, SO_2 and air temperature.

Conclusion: Machine learning algorithm can predict number of patients due to conjunctivitis, among which, the Random forest algorithm had the highest accuracy. Machine learning algorithm could provide accurate information for hospitals dealing with conjunctivitis caused by air factors.

Introduction

Ambient air pollution is an important public health problem, especially in developing countries. It is considered an important risk factor for morbidity and mortality worldwide [1, 2]. The World Health Organization (WHO) estimated that seven million people died from exposure to air pollution in 2012 [3]. China is now facing one of the worst air pollution problems, being the largest developing nation in the world [4]. We have previously reported that air quality improvement could reduce the number of hospital admissions due to acute myocardial infarction [5]. It was also reported that air pollution has short-term and lagging effects on lungs function in school-age children in the city of Hangzhou, in Zhejiang province, China [6].

The conjunctiva lines the inside of the eyelids and covers the sclera. It is composed of non-keratinized stratified squamous and columnar epithelium, along with interspersed goblet cells. The conjunctiva is vulnerable to various harmful factors, such as bacteria, viruses, allergens, and chemicals, because it is

highly vascularized and constantly exposed to the external environment, the result is often conjunctivitis. Conjunctivitis, as the most common eye disease diagnosed in emergency departments, is a cause for serious health and economic burden worldwide. A meta-analysis study has shown that females and youth are more vulnerable to PM2.5, NO_2, and O_3 [7]. Several studies have further provided significant evidence that, in China, patient due to conjunctivitis are associated with air pollution [8–10].

A constantly increasing number of studies attempt to adopt machine learning method to predict clinical events. For example, studies have used machine learning algorithm to predict the probability of prostate cancer [11], atrial fibrillation in primary care [12], remission after transsphenoidal surgery among patients with acromegaly [13], and pulmonary hypertension [14]. To date, there has been no published attempt to describe what kind of information is to be used in machine learning (ML) training to predict number patient due to conjunctivitis.

Our study assessed the performance of machine learning algorithms to predict number of patients due to conjunctivitis, and the optimal machine learning prediction model was selected to predict. Through the establishment of the best method to predict the number of patients in a period of time in the future. It can provide accurate information decision for the hospital.

Methods

Study population

Data about Confirmed case related to conjunctivitis between August 1, 2014 to August 1, 2019 were obtained from the Eye Center of the hospital, one of the largest ophthalmology clinics in Zhejiang Province. Collected data included visit's date, gender, age, home address, and whether the visit was the patient's first visit or a re-visit. The International Classification of Diseases (10th Revision, including H10.901, H10.301, and H10.402) was used to diagnose conjunctivitis.

Air Pollution And Weather Data

There are six air quality monitoring stations in Hangzhou, providing daily values of PM2.5 (μg/m^3), PM10 (μg/m^3), SO_2 (μg/m^3), O_3 (μg/m^3), NO_2 (μg/m^3), CO (μg/m^3), and highest and lowest air temperatures. Hangzhou daily air pollution parameters and temperatures, between January 1, 2014 and August 31, 2019, were downloaded from the China Meteorological Administration (http://data.cma.cn/). After calculating the hourly average pollutant concentration of the six stations, the 24-hour average pollutant concentration was calculated. Severity of the pollution was assigned one of four quartiles of the Air Quality Index (AQI). The AQI was calculated based on the six air indicators mentioned above. When AQI < 100, it means no pollution, 101 < AQI < 150 means mild pollution, 151 < AQI < 200 means moderate pollution, and AQI > 201 means severe pollution (USEPA, https://www.gpo.gov/).Considering the effect of air pollution on conjunctivitis, environmental and weather parameters were calculated as the average of

the previous three days. For example, the value of the environmental factors for January 4, 2015 was calculated as the average of the values of January 1–3, 2015.The details of the Affiliated Hospital of Hangzhou Normal University and the specific air quality monitoring stations are showed in appendix Fig. 1.

Machine Learning Algorithms

We mainly consider the multiple regression model:$Y = X^T\beta, X = (x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8)$ where$y$ representing number of patients, $x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8$respectively representing environmental variables: CO, O_3, NO_2, SO_2, PM10, PM2.5, highest and lowest temperature. Seven typical machine learning algorithms to train the regression model. These included Lasso penalized liner model [15], Decision tree [16], Boosting regression [17], Bagging regression [18], Random forest [19], Support vector [20], and Artificial neural network [21]. In each method, the reliability of the result was judged, using the 10-fold cross validation. Data were randomly split into test and training datasets at a ratio of 3:7, respectively. Machine learning models were built on the training dataset, prediction results using test sets. Figure 1 shows the flowchart of this work. Machine learning techniques were implemented in R using the package for lars, rpart.plot, mboost ,ipred, randomForest, rminer and nnet respectively.

Statistical Analyses

To evaluate the accuracy of the seven methods, normalized mean square error (NMSE) was used to assess the accuracy of each method:

$$NMSE = \frac{\overline{(y - \hat{y})^2}}{\overline{(y - \bar{y})^2}},$$

where y represents the actual number of patients per day, $\bar{y}$ represents the actual average number of patients per day, and $\hat{y}$ is the predicted number of patients for the test set based on the model from the training set. The NMSE range is 0–1,the smaller the value, the higher the accuracy. At the same time, the prediction deviation distribution is compared, the deviation is equal to the absolute value of the predicted value minus the real value. Pearson correlation coefficient was used to compare between the predicted number of patients and the true number of patients in the selected optimal ML method. The importance of the variable was measured in terms of the average NMSE decline caused by the deletion of the variable, the larger the value, the more important for variable. The overall predictive number of cases for all patients were calculated using all seven methods. Further analyses by sex (men and women) and age at diagnosis (<45, 45–60, >60 years) were also conducted. All analyses were performed using the statistical programming environment R (version 3.6.0).

Results

Characteristics of patient and environmental variables

A total of 84977 patients, living in the air-monitored area of Hangzhou city, were included in this study. Table 1 describes the baseline characteristics of patient and environmental variables .The average number of patients with conjunctivitis per day was 54. Between August 1, 2014 and August 1, 2019, the trend of AQI daily change was very similar to that of the daily number of patients with conjunctivitis at the hospital (Fig. 2).

Table 1
baseline patient and environmental characteristics

Characteristics	Data
Patient:	84977(47789,37188)
Gender, n (male,female)	42.1 ± 9.4
Age(mean ± SD)	
Air pollution and weather:	89.6 ± 24.5
AQI(mean ± SD)	44.8 ± 21.4
PM2.5(mean ± SD)	72.4 ± 30.5
PM10 (mean ± SD)	10.9 ± 5.4
SO_2(mean ± SD)	40.9 ± 13.7
NO_2(mean ± SD)	0.85 ± 0.18
CO(mean ± SD)	56.1 ± 24.2
O_3(mean ± SD)	22.3 ± 8.8
highest temperature(mean ± SD)	14.6 ± 8.4
lowest temperature (mean ± SD)	

Predictive Performance Of The Seven Machine Learning Algorithms

The estimates of accuracy derived from each ML method are presented in Table 2. When all patients were evaluated together, the NMSE for Random forest was 0.392. When evaluated by gender, the NMSE of Random forest males and females was 0.443 and 0.433, respectively. After stratifying by age, the NMSE of Random forest for the age groups (□45, 45–60, □60 years) was 0.406, 0.529, and 0.626, respectively. The order of NMSE obtained by the different ML methods, from small to large, was as follows: Random forest, Neural network, Bagging regression, Boosting regression, Support vectors, Decision tree, and Lasso

penalized linear regression. Further analyses, after stratifying by sex and age at diagnosis, also showed Random forest to be superior to the others ML methods.

Table 2
Normalized mean square error (NMSE) predictive accuracy of the different machine learning methods, stratified by gender and age

Method	NMSE					
	All	Males	Females	45	45–60	60
Lasso penalized liner regression	0.755	0.769	0.763	0.755	0.839	0.870
Decision tree	0.710	0.717	0.718	0.725	0.767	0.770
Boosting regression	0.616	0.623	0.634	0.597	0.850	0.743
Bagging regression	0.615	0.641	0.651	0.625	0.808	0.818
Random forest	**0.392**	**0.443**	**0.433**	**0.406**	**0.529**	**0.626**
Support vector	0.688	0.703	0.606	0.652	0.612	0.665
Neural network	0.476	0.490	0.555	0.557	0.603	0.633

By comparing the deviation of distribution obtained by the different methods, when all patients were considered, the deviation obtained using Random forest algorithm was the smallest, followed by Neural network algorithm. Deviation obtained when using the Lasso penalized liner model was the largest (Fig. 3A). The same result was obtained after stratifying for gender and age (Fig. 3B-F).

Prediction Accuracy Based On Random Forest

Furthermore, we explored the correlation between the predicted number of patients and true number of patients based on the Random forest algorithm, and found it to be 0.969 for all patients (Fig. 4A), 0.968 for men (Fig. 4B), and 0.967 for women (Fig. 4C). The values for the different age groups were 0.968 for < 45 years (Fig. 4D), 0.931 for patients between 45–60 years (Fig. 4E), and 0.966 for patients > 60 years (Fig. 4F).

Importance Of Significant Environmental Variables

We further analyzed the classifier-specific predictive importance of eight environmental variables based on the Random forest algorithm. The number of patients could be predicted in decreasing order of accuracy, when all patients were considered, by O_3 > SO_2 > CO > Lowest temperature > Highest temperature > NO_2 > PM10 > PM2.5 (Fig. 5A). When stratified by gender, the three most important variables that affected the accuracy of predicting the number of men with the disease were PM10, O_3, and highest temperature, while for women these were highest temperature, SO_2, and Lowest temperature

(Fig. 5B). When stratified by age, the most important variables that affect the accuracy of predicting the number of patients aged 45 and aged 45–60 were SO_2 and PM10, respectively. PM10 and PM2.5 were the two most important variables that affect the predictive accuracy for the number of patients aged 60 (Fig. 5C).

Discussion

In the present study, based on 5 years of big data on confirmed cases of conjunctivitis, we developed and validated seven ML models to predict number of patients for conjunctivitis. The results show that Random forest algorithm had the highest predictive accuracy, and that the most important variables affecting the predictive accuracy were O_3, CO, SO_2, NO_2, and air temperature. This provides an accurate method for future use of air pollution to predict the number of patients.

Air pollution, caused by multiple-pollutant emissions and vehicle exhaust, has been aggravating yearly with the rapid urbanization and development of transportation infrastructure in China [22]. Accumulating evidence indicates that air pollution has a significant impact on cardiovascular- and chronic cardiopulmonary-related morbidity and mortality in China [23–25]. Some studies have also shown that air pollution is associated with eye diseases, such as dry eye [26–27], conjunctivitis and pterygium [28]. To the best of our knowledge, there has been no large-scale study to assess the usefulness of ML in predicting patient due to conjunctivitis based on air pollution. There is, therefore, an urgent need for an effective and widely applicable predictive method.

Using conjunctivitis-related clinic cases and air pollution data for ML, we were able to predict the number of possible conjunctivitis clinic cases, and used this information to improve work planning and new initiatives. This predictive approach and the yearly fluctuations in patients with conjunctivitis, could act as the scientific basis for hospital management decision making on matters like shunting of hospital management and medical staff as needed and relieve pressure off consultations in an orderly manner. Our results, corroborate with previous studies, showing that the same important environmental factors affecting conjunctivitis [9]. Stratified analyses by gender and age showed that environmental factors affect the number of conjunctivitis patients differed between genders and age groups.

With the arrival of medical big data, the traditional statistical methods were unable to adapt analysis to the huge data generated. A continuously increasing number of researchers began using ML algorithms to analyze data. In this study, we used seven ML methods to predict the number of patients with conjunctivitis. The linear model was built with the adaptive lasso penalty and partial least square regression. As such, it can be used in future research [29]. Decision tree is a weak machine learning method, but when many decision trees are combined together, they can form a strong machine learning method, such as Boosting, Bagging, and Random forest. Random forest is composed of decision trees, formed by randomly putting back samples, so that the accuracy is greatly improved[19]. Our results show that the predictive accuracy of the Random forest method is better than that of Decision tree, Bagging, and Boosting. Artificial neural network is an imitation of the natural neural network, which can effectively

solve complex regression and classification problems with a large number of related variables. Generally, if the training data set is very large, the predictive accuracy is high [30]. In this study, the predictive accuracy, based on the artificial neural network, was lower than that of Random forest. This may be related to the training data set size.

Conclusion

For the first time, machine learning was used to predict the number of conjunctivitis-related patients to the hospital based on past conjunctivitis-related patient visits and air pollution parameters. In view of the number of patients with conjunctivitis in Hangzhou, China, we have established seven ML algorithms, and identified the best prediction method. In the future, we will expand our research scope to include the use of machine learning to explore other diseases related to air pollution.

Abbreviations

AQI
Air Quality Index
ML
machine learning
NMSE
normalized mean square error

Declarations

Ethics approval and consent to participate

Not applicable. My manuscript does not report on or involve the use of any animal or human data or tissue.

Consent for publication

Not applicable. All data were supplied and analysed in an anonymous format, without access to personal identifying information.

Conflict of interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Availability of data and materials

The datasets used in the current study are available from the corresponding author on reasonable request.

Funding

Hangzhou Science and Technology Bureau Fund (No.20120533Q32; No.20150733Q24; No.20191203B96; No.20191203B105; No.20171334M01); Youth fund of Zhejiang Academy of Medical Sciences (No.2019Y009)

Contributions

Juan Chen and Yong-ran Cheng: data curation, software, methodology. Meng-Yun Zhou: methodology, writing. Nan Wang: methodology, writing. Lan Ye: data curation, writing. Ming-Wei Wang, Zhan-hui Feng and Jin Yao: methodology, supervision, project administration, writing. The authors read and approved the final manuscript.

Acknowledgment

Thanks to all the ophthalmologists of Hangzhou Normal University for their help. The presented study was supported by the Hangzhou Science and Technology Bureau Fund (No.20120533Q32; No.20150733Q24; No.20191203B96; No.20191203B105; No.20171334M01); Youth fund of Zhejiang Academy of Medical Sciences (No.2019Y009)

References

1. Cohen AJ, Brauer M, Burnett R. Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015. Lancet. 2018;391(10130):1576-.
2. Mannucci PM, Franchini M. Health Effects of Ambient Air Pollution in Developing Countries. Int J Env Res Pub He. 2017;14(9).
3. Gulland A. One in eight deaths is due to air pollution, says WHO. Bmj-Brit Med J. 2014;348.
4. Niu Y, Chen R, Kan H. Air Pollution, Disease Burden, and Health Economic Loss in China. Adv Exp Med Biol. 2017;1017:233–42.
5. Wang MW, Chen J, Cai R. Air quality and acute myocardial infarction in adults during the 2016 Hangzhou G20 summit. Environ Sci Pollut Res Int. 2018;25(10):9949–56.
6. Liu WY, Zhang L, Xu H, Xu SS, Lyu Y, Zhang WH, et al. Short-term effects of air pollution on lung function of school-age children in Hangzhou. Zhonghua Yu Fang Yi Xue Za Zhi. 2019;53(6):614–8.

7. Chen R, Yang J, Zhang C, Li B, Bergmann S, Zeng F, et al. Global Associations of Air Pollution and Conjunctivitis Diseases: A Systematic Review and Meta-Analysis. Int J Environ Res Public Health. 2019;16:19.
8. Li Z, Bian X, Yin J, Zhang X, Mu G. The Effect of Air Pollution on the Occurrence of Nonspecific Conjunctivitis. J Ophthalmol. 2016;2016:3628762.
9. Fu Q, Mo Z, Lyu D, Zhang L, Qin Z, Tang Q, et al. Air pollution and outpatient visits for conjunctivitis: A case-crossover study in Hangzhou, China. Environ Pollut. 2017;231(Pt 2):1344–50.
10. Lu P, Zhang Y, Xia G, Zhang W, Li S, Guo Y. Short-term exposure to air pollution and conjunctivitis outpatient visits: A multi-city study in China. Environ Pollut. 2019;254(Pt A):113030..
11. Jovic S, Miljkovic M, Ivanovic M, Saranovic M, Arsic M. Prostate Cancer Probability Prediction By Machine Learning Technique. Cancer Invest. 2017;35(10):647–51.
12. Hill NR, Ayoubkhani D, McEwan P, Sugrue DM, Farooqui U, Lister S, et al. Predicting atrial fibrillation in primary care using machine learning. PLoS One. 2019;14(11):e0224582.
13. Fan YH, Li YS, Li YC, Feng SS, Bao XJ, Feng M, et al. Development and assessment of machine learning algorithms for predicting remission after transsphenoidal surgery among patients with acromegaly. Endocrine. 2019.
14. Leha A, Hellenkamp K, Unsold B, Mushemi-Blake S, Shah AM, Hasenfuss G, et al. A machine learning approach for the prediction of pulmonary hypertension. PLoS One. 2019;14(10):e0224453.
15. Tibshirani R. Regression Shrinkage and Selection Via the Lasso Journal of the Royal Statistical Society Series B (Methodological). 1996:267–88.
16. Prosperi MC, Belgrave D, Buchan I, Simpson A, Custovic A. Challenges in interpreting allergen microarrays in relation to clinical symptoms: a machine learning approach. Pediatr Allergy Immunol. 2014;25(1):71–9.
17. Salzberg SL. C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc. 1993. Machine Learning. 1993;16(3):235 – 40.
18. Hothorn T, Buehlmann P, Kneib T, mboost. Model-Based Boosting Journal of Machine Learning Research. 2017;11(2):11(2).
19. Liaw A, Wiener M, ,Liaw A. Classification and Regression by Random Forests. R News. 2002;23(23):233–42.
20. Cortes C. Vapnik V ,VV. Support-vector networks. Mach Learn. 1995;20(3)::273–97.
21. Active and Adaptive Vision
Fukushima K.Active and Adaptive Vision: Neural Network Models. EEE International Workshop on Biologically Motivated Computer Vision Springer-Verlag. 2000.
22. Huang C, Wang Q, Wang S, Ren M, Ma R, He Y. Air Pollution Prevention and Control Policy in China. Adv Exp Med Biol. 2017;1017:243–61.
23. Madaniyazi L, Nagashima T, Guo Y, Yu W, Tong S. Projecting Fine Particulate Matter-Related Mortality in East China. Environ Sci Technol. 2015;49(18):11141–50.

24. Lu X, Lin C, Li Y, Yao T, Fung JC, Lau AK. Assessment of health burden caused by particulate matter in southern China using high-resolution satellite observation. Environ Int. 2017;98:160–70.
25. Li P, Xin J, Wang Y, Li G, Pan X, Wang S, et al. Association between particulate matter and its chemical constituents of urban air pollution and daily mortality or morbidity in Beijing City. Environ Sci Pollut Res Int. 2015;22(1):358–68.
26. Mo Z, Fu Q, Lyu D, Zhang L, Qin Z, Tang Q, et al. Impacts of air pollution on dry eye disease among residents in Hangzhou, China: A case-crossover study. Environ Pollut. 2019;246:183–9.
27. Hwang SH, Choi YH, Paik HJ, Wee WR, Kim MK, Kim DH. Potential Importance of Ozone in the Association Between Outdoor Air Pollution and Dry Eye Disease in South Korea. JAMA Ophthalmol. 2016;134(5):503–10.
28. Lee KW, Choi YH, Hwang SH, Paik HJ, Kim MK, Wee WR, et al. Outdoor Air Pollution and Pterygium in Korea. J Korean Med Sci. 2017;32(1):143–50.
29. Nguyen HT, Lee BW. Assessment of rice leaf growth and nitrogen status by hyperspectral canopy reflectance and partial least square regression. Eur J Agron. 2006;24(4):349–56.
30. Pradhan B, Lee S. Landslide risk analysis using artificial neural network model focusing on different training sites. International journal of physical sciences. 2009;4(2):1–15.

Figures

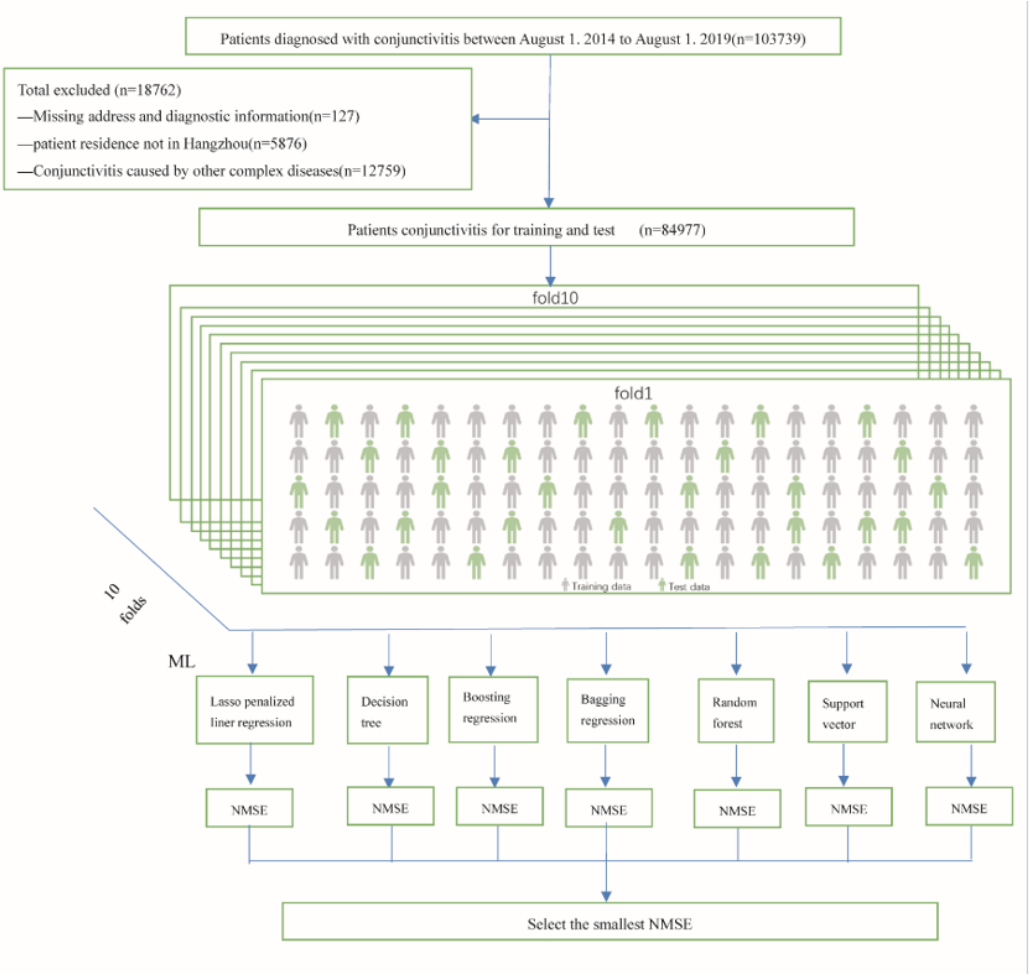

Figure 1

The flowchart of model developing and validation.(Lasso penalized liner model, Decision tree, Boosting regression, Bagging regression, Random forest, Support vector, and Neural network)

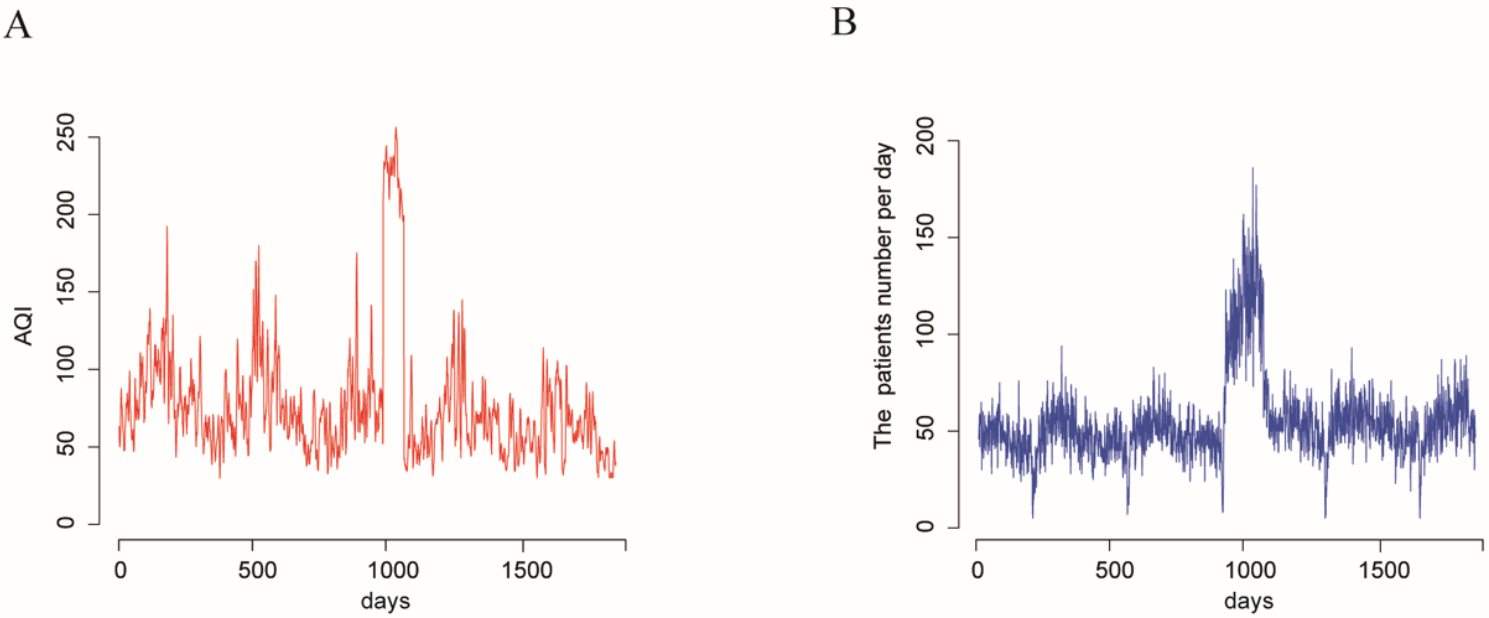

Figure 2

Comparison of the deviation distributions obtained by different methods for all patients (A), males (B), females (C), age 45 (D), age between 45-60 (E), and age 60 (F)

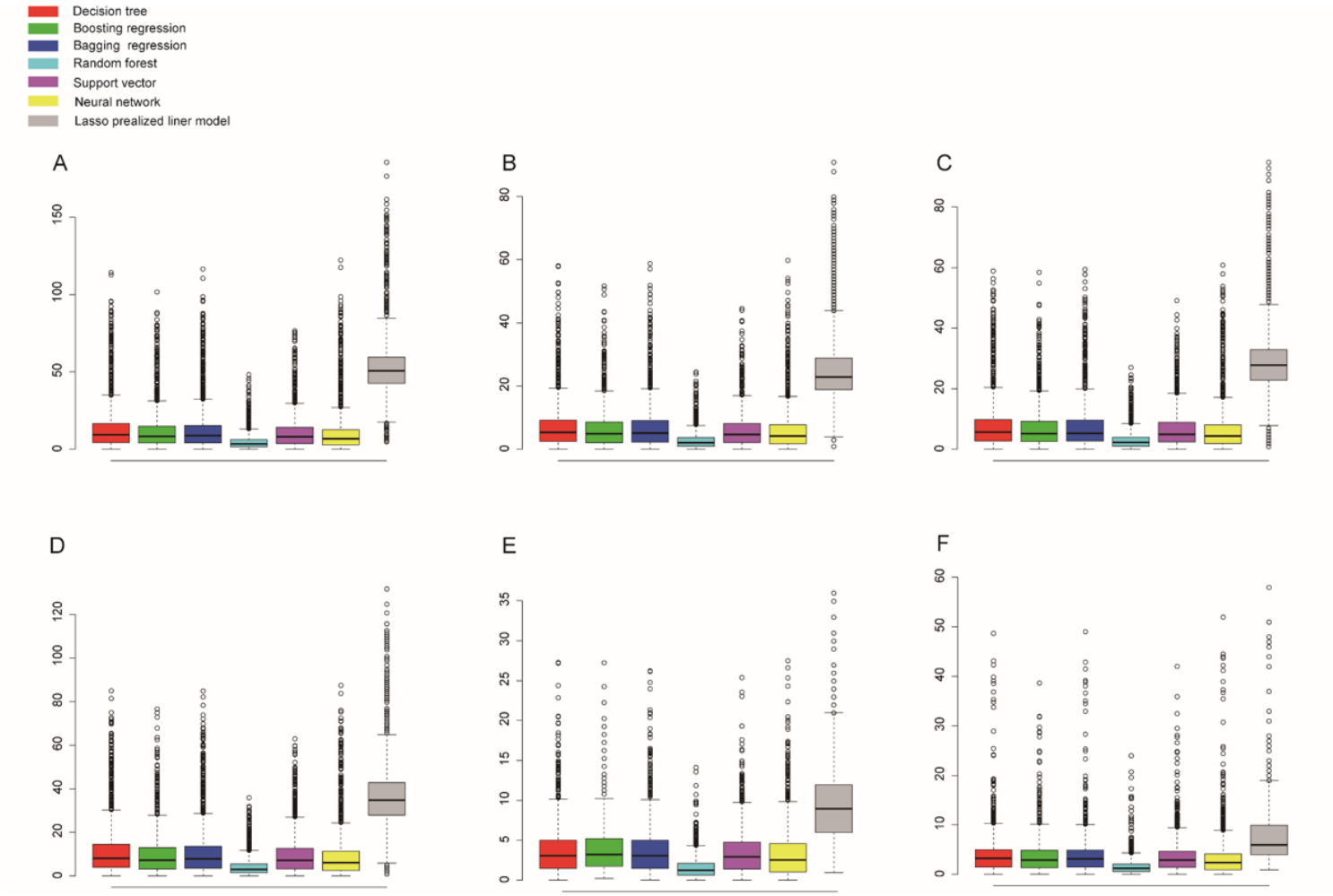

Figure 3

Comparison of the deviation distributions obtained by different methods for all patients (A), males (B), females (C), age 45 (D), age between 45-60 (E), and age 60 (F)

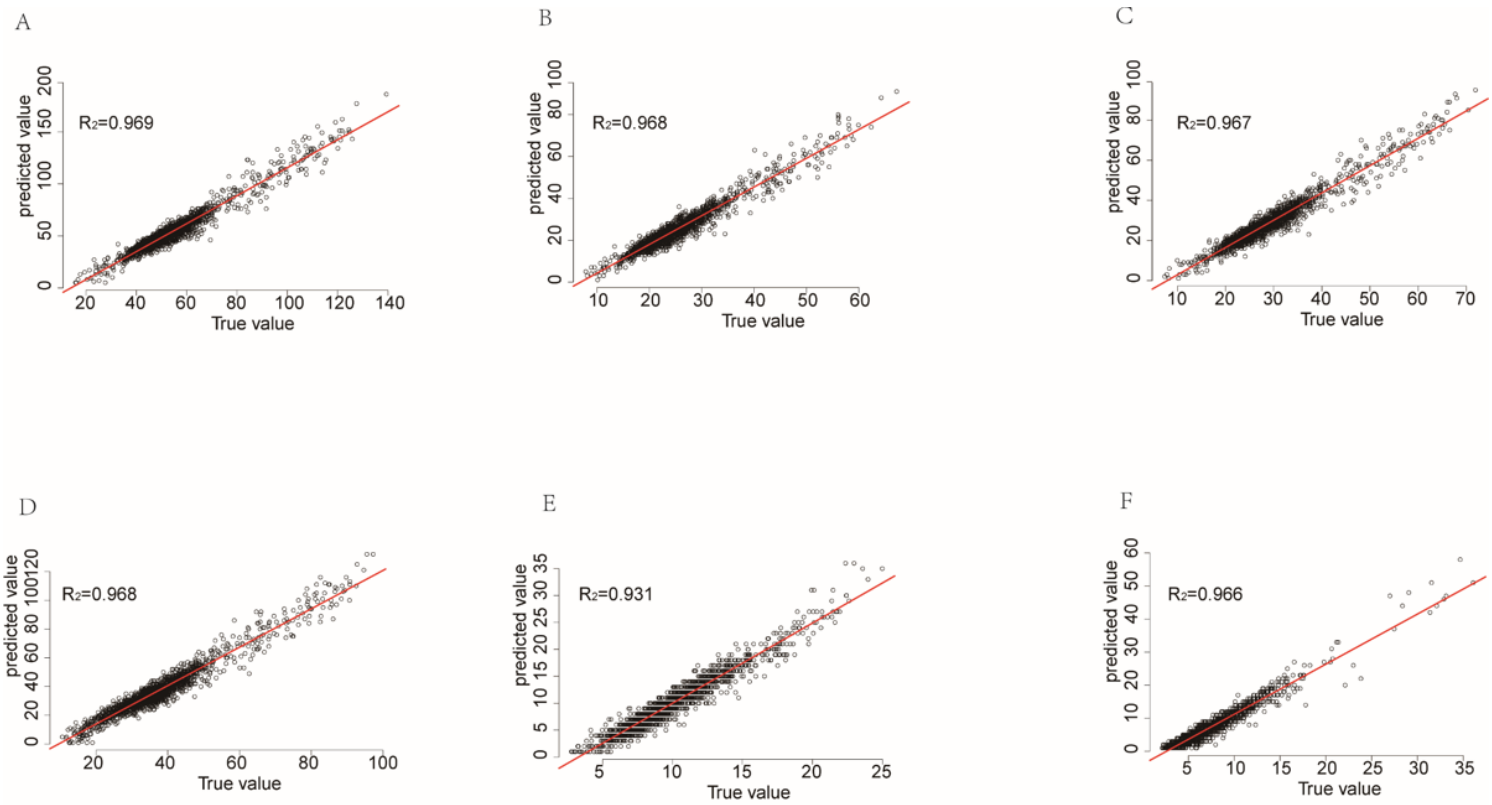

Figure 4

Pearson correlation coefficient analysis associating between the predicted value(Predicted number of patients) and true value(True number of patients) based on the Random forest method, for all patients (A), males (B), females (C), age □ 45 (D), age between 45-60 (E), age □ 60(F).

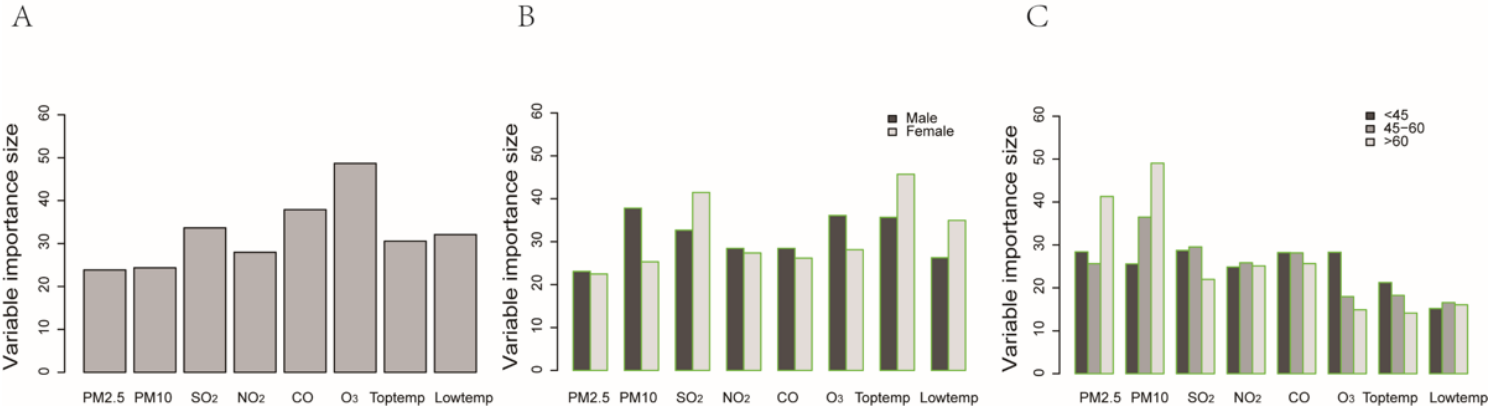

Figure 5

Importance ranking of the variables based on the Random forest predictions. All patients(A), gender (B), age (C). Toptemp: Daily highest temperature, Lowtemp: Daily lowest temperature.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AppendixS1.tif