Predictions of [illegible] Based Respiratory Hospital Visits: Developing and Validating a Machine Learning Model with a Novel Composite Air Pollution Index

The multi-pollutant respiratory outpatient prediction framework.

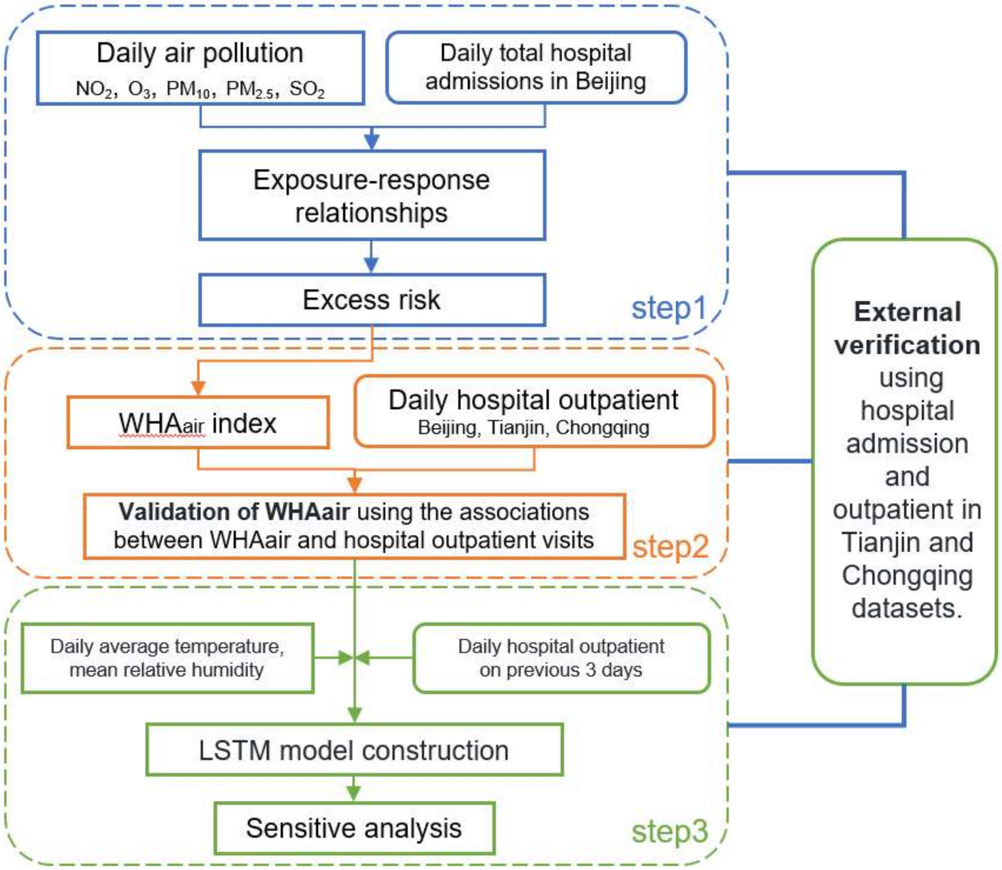

Figure S1. R rch flowchart

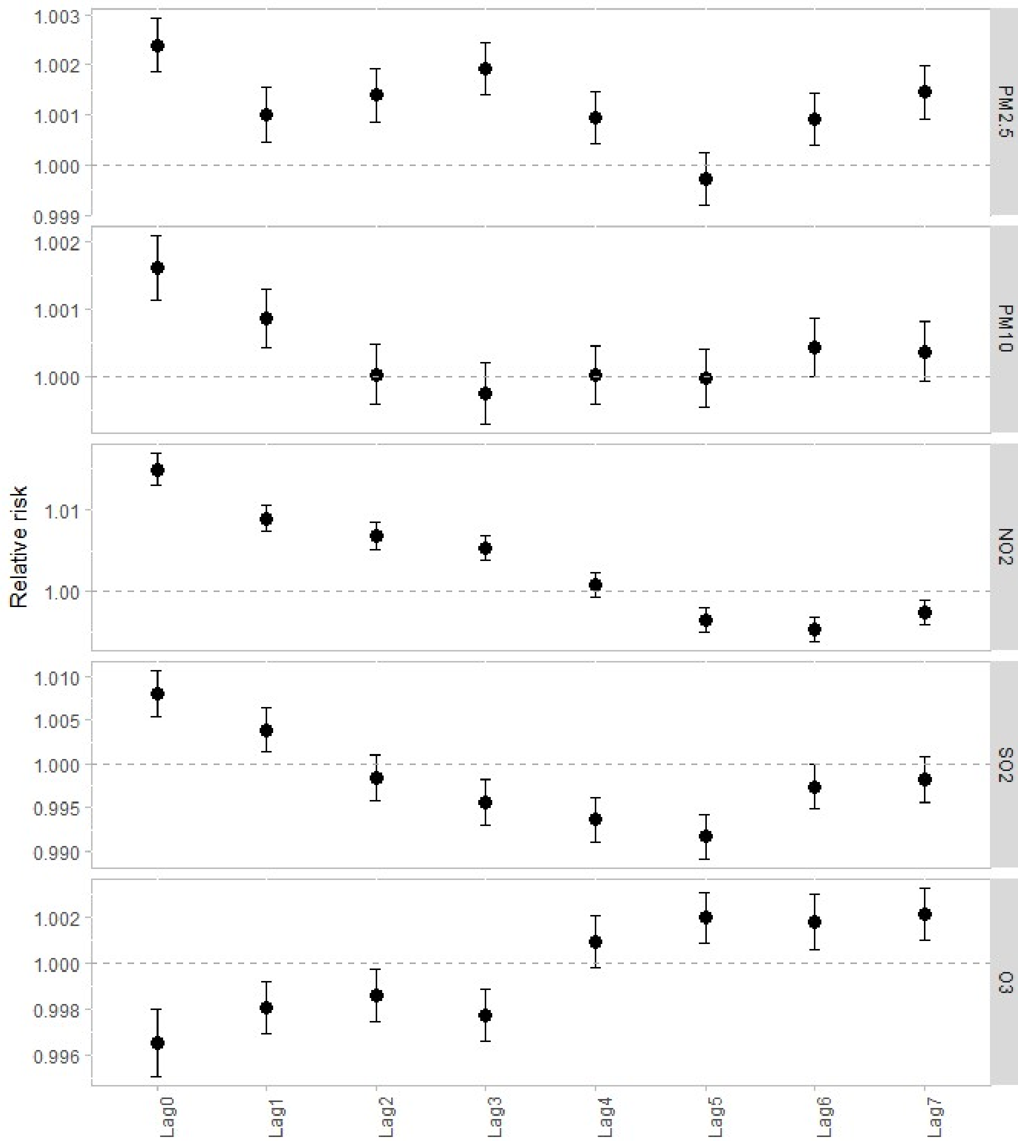

Figure S2. Lag patterns of the associations between five air pollutants and total non-accidental hospital admissions

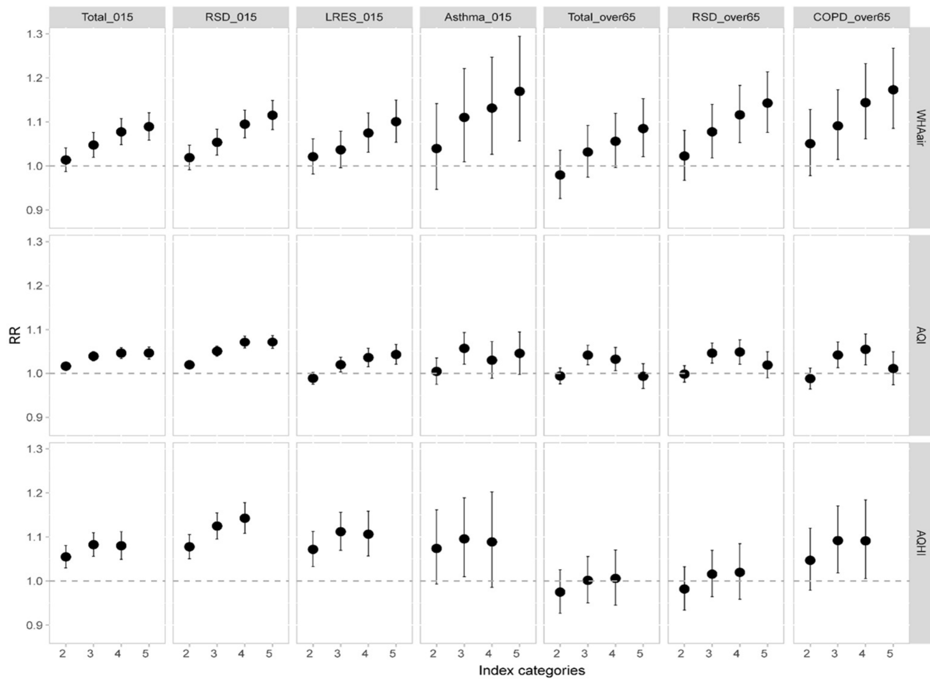

Figure S3. Associations between main causes of hospital outpatient visits for vulnerable population and each level of WHA_{air}, AQI and AQHI. Total_015: [illegible]accidental visit risk of population aged ≤15years; RSD_015: respiratory system disease of population aged ≤15years; LRES_015: lower respiratory system disease of population aged ≤15years; Asthma_015: Asthma disease of population aged ≤15years; Total_over65: non-accidental visit risk of population aged ≥65 years; RSD_over65: respiratory system disease of population aged ≥65 years; COPD_over65: chronic obstructive pulmonary disease of population aged ≥65 years.

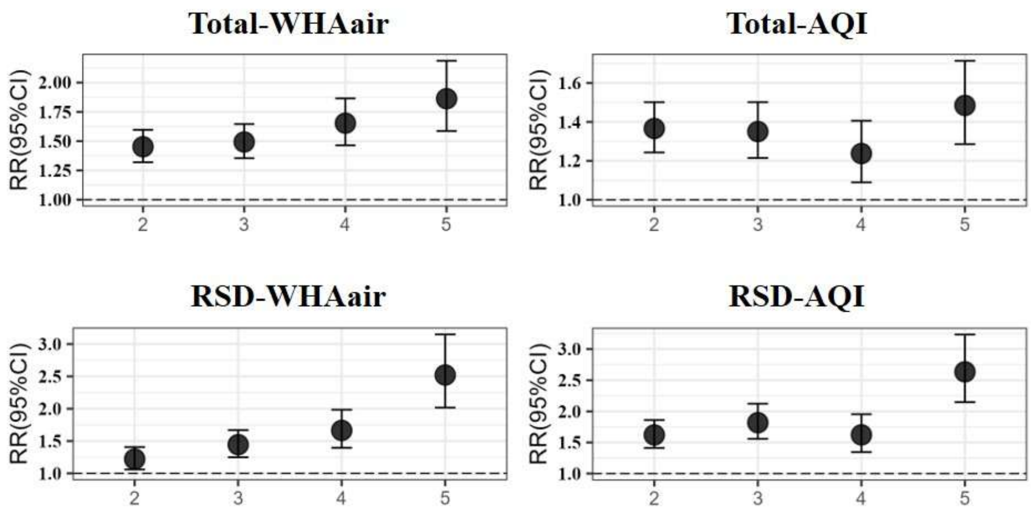

Figure S4. Associations between hospital outpatient visits and each level of WHA_{air}, AQI in Tianjin, 2013-2019.

Total-WHA_{air}: non-accidental visit risk with each level of WHA_{air}; Total-AQI: non-accidental visit risk with each level of AQI; RSD-WHA_{air}: respiratory system disease with each level of WHA_{air}; RSD-AQI: respiratory system disease with each level of AQI.

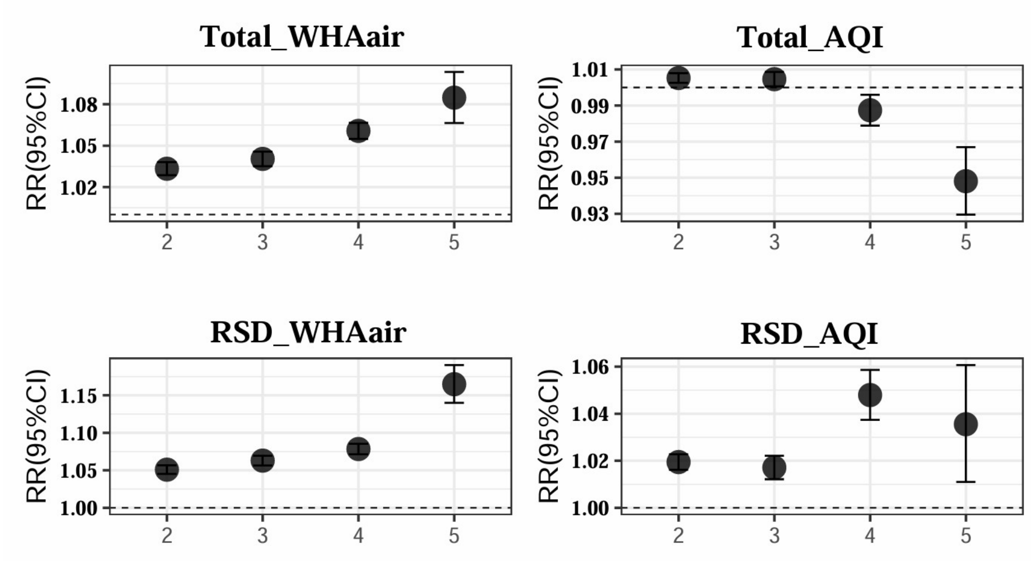

Figure S5. Associations between hospital outpatient visits and each level of WHA_{air}, AQI in Chongqing, 2018-2023.

Total-WHA_{air}: non-accidental visit risk with each level of WHA_{air}; Total-AQI: non-accidental visit risk with each level of AQI; RSD-WHA_{air}: respiratory system disease with each level of WHA_{air}; RSD-AQI: respiratory system disease with each level of AQI.

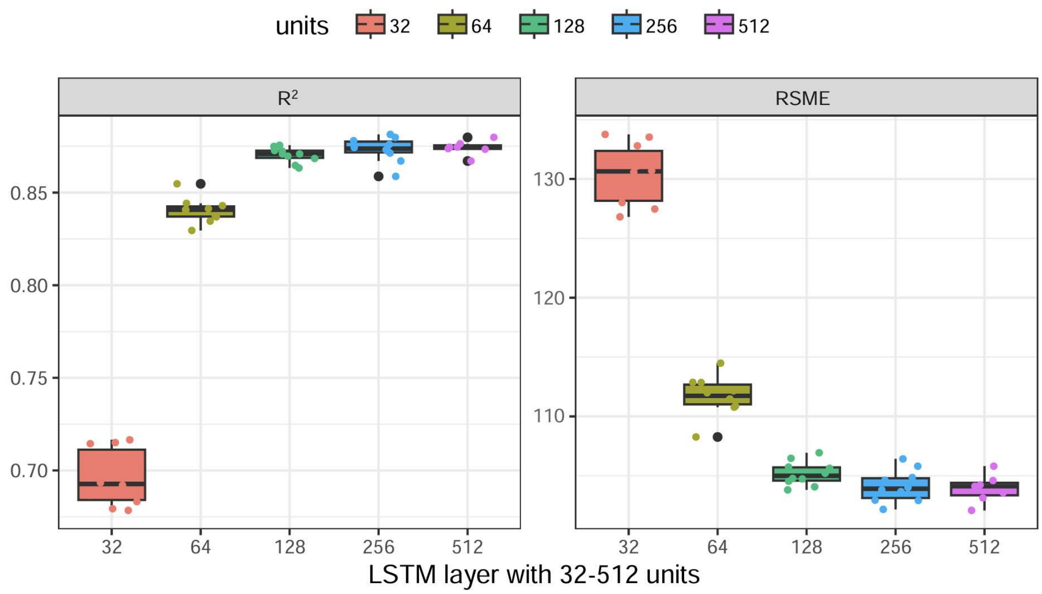

Figure S6. Box Plots of R^2 and RMSE for Different LSTM Units Models on Three Datasets.

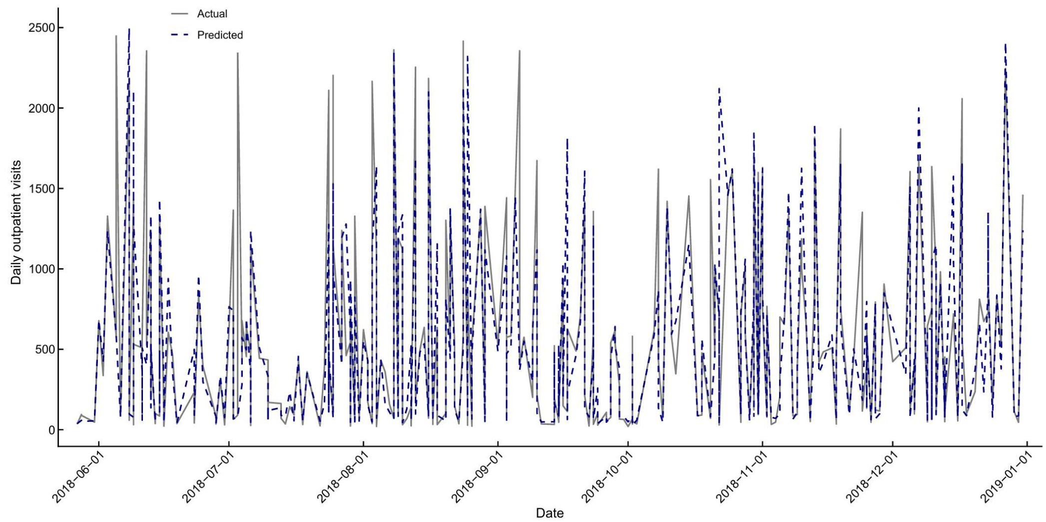

Figure S7. Comparison of Actual and Predicted Values.

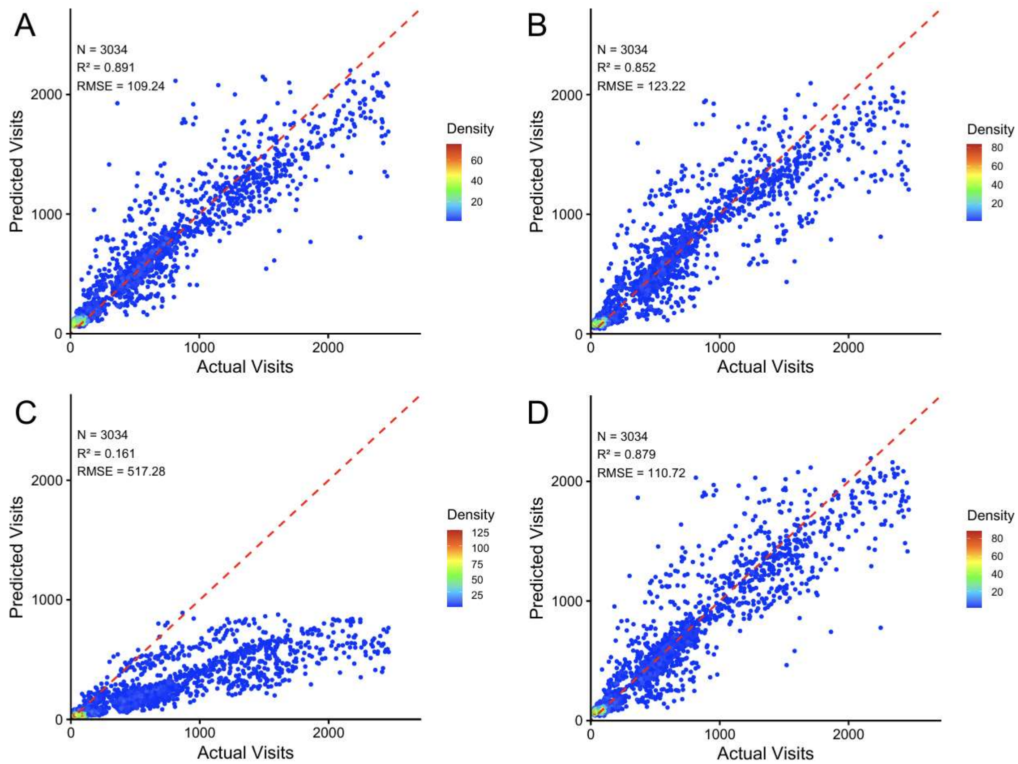

Figure S8. Actual vs. Predicted Scatterplot for Model Comparison of the Time Division Data Model in Beijing. (A) All-features model. (B) WHA_{air}-excluded model. (C) Outpatient-visits-excluded model. (D) Temp-Humidity-excluded model.

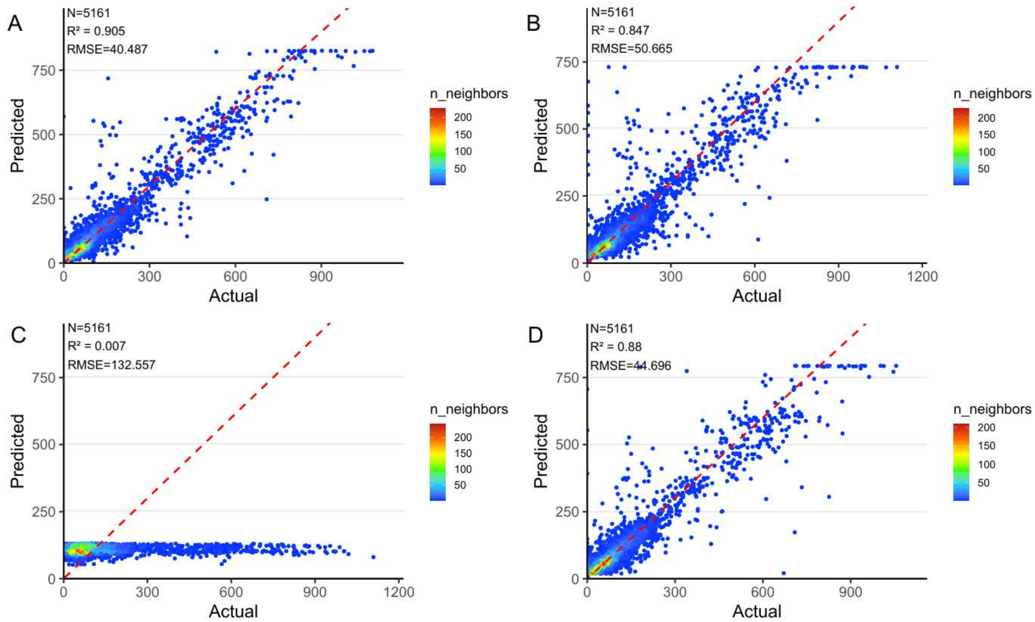

Figure S9. Actual vs Predicted Scatterplot for Model Comparison in Tianjin data. (A) All-features model. (B) WHA_{air}-excluded model. (C) Outpatient-visits-excluded model. (D) Temp-Humidity-excluded model.

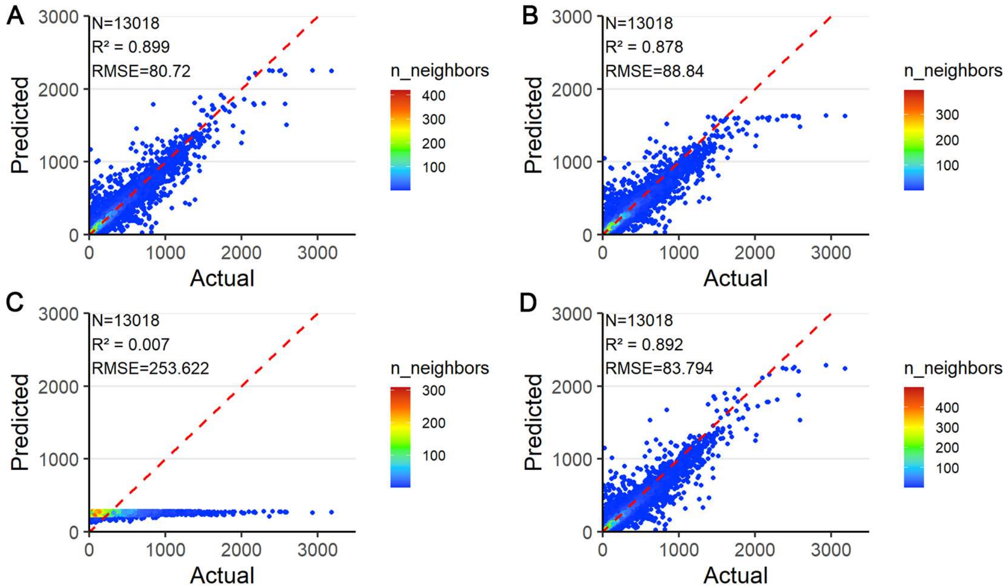

Figure S10. Actual vs Predicted Scatterplot for Model Comparison in Chongqing data. (A) All-features model. (B) WHA_{air}-excluded model. (C) Outpatient-visits-excluded model. (D) Temp-Humidity-excluded model.

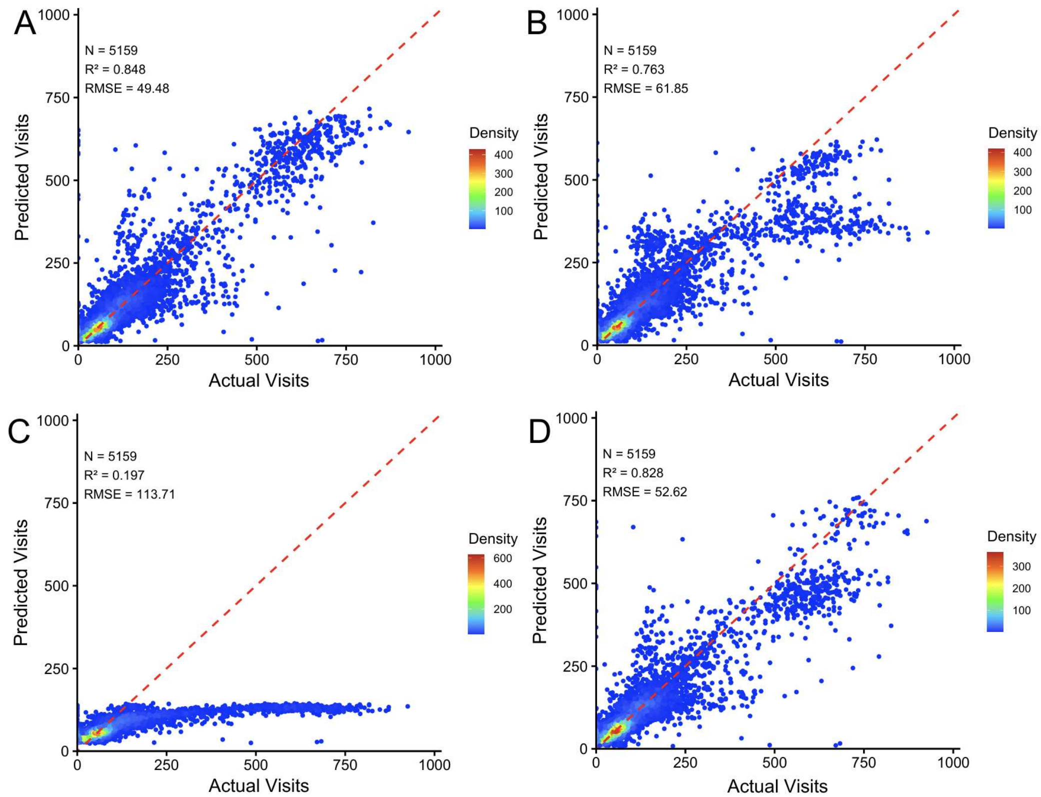

Figure S11. Actual vs. Predicted Scatterplot for Model Comparison of the Time Division Data Model in Tianjin. (A) All-features model. (B) WHA_{air}-excluded model. (C) Outpatient-visits-excluded model. (D) Temp-Humidity-excluded model.

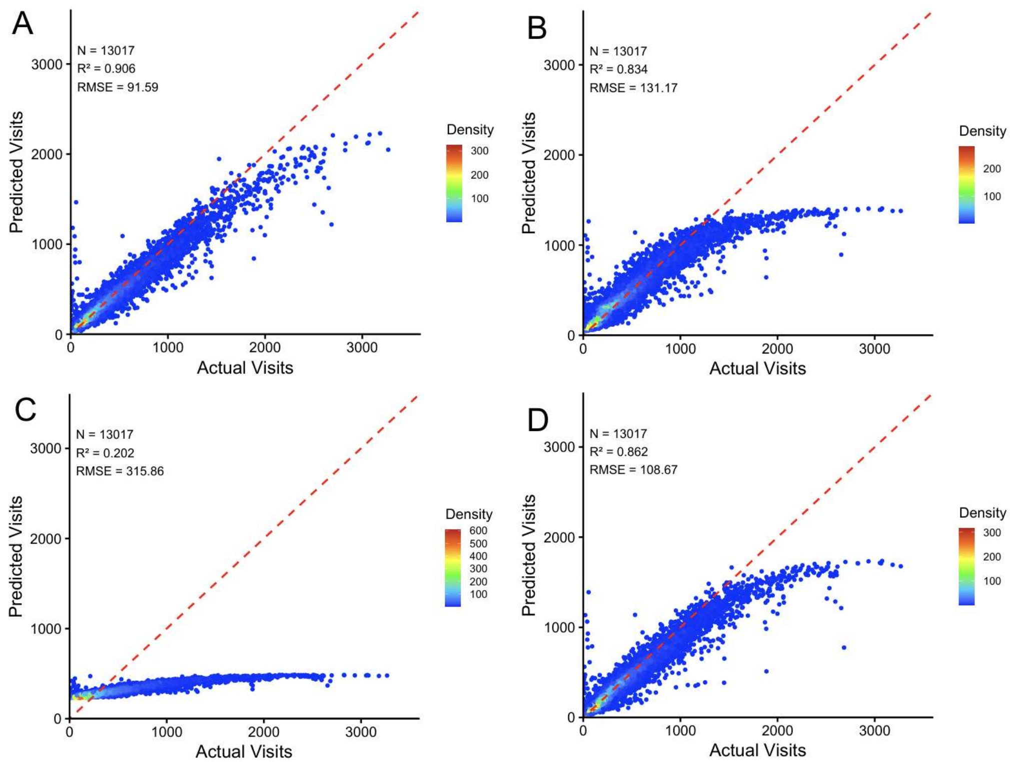

Figure S12. Actual vs. Predicted Scatterplot for Model Comparison of the Time Division Data Model in Chongqing. (A) All-features model. (B) WHA_{air}-excluded model. (C) Outpatient-visits-excluded model. (D) Temp-Humidity-excluded model.

Table S1. Summary statistics of environmental exposure and hospital admissions in Beijing, 2013-2018

	N	Mean	SD	Min	Q25	Q75	Max
Air pollution exposure (μg/m^3)							
$PM_{2.5}$		68.4	62.5	3	24.8	91.4	520.6
PM_{10}		76.1	66.7	3	30.5	103.2	1196.6
NO_2		43.6	26.2	0.8	24.9	56.8	193.6
SO_2		12.4	15.8	2.9	3.3	14.1	182.6
O_3		56.3	38.3	2.1	26.2	79.5	211.4
Meteorological factors							
Relative humidity (%)		54	19.6	10.2	38.1	70.2	98.9
Temperature (℃)		11.6	11.5	-21.9	0.5	22.2	34.4
Precipitation (mm)		0.45	2	0	0	0.11	78.5
Cause-specific hospital admission							
Non-accidental deaths (Total)	12951191	369	452	0	97	402	35[illegible]
Respiratory (RES)	1292340	36	34	0	14	47	254

Table S2. Sensitivity analysis on the exposure-response relationships between five air pollutants and total hospital admissions using two different models

Air pollutants	GNM	GLM
PM_{10}	1.0019(1.0013, 1.0025)	1.0018(1.0013, 1.0023)
$PM_{2.5}$	1.003(1.0023, 1.0037)	1.002(1.0015, 1.0026)*
O_3	1.0065(1.0048, 1.0082)	1.0022(1.001, 1.0033)**
NO_2	1.0152(1.0133, 1.017)	1.0134(1.0116, 1.0151)
SO_2	1.0046(1.0013, 1.0079)	1.0054(1.0028, 1.008)

**$p<0.01$, *$p<0.05$ for the significance test

Table S3. Threshold values of excess risk (ER) for the five-level WHA_{air} classification

WHA_{air}*	Beijing	Tianjin	Chongqing
2	5.8	6.4	9.2
3	11.5	9.4	13.4
4	17.0	13.7	18.0
5	21.8	29.7	31.1

* Level 1: ER < Level 2 threshold. Thresholds derived via automated iteration from local exposure–response relationships for general, pediatric, and elderly populations (Method S2).

Table S4. Frequency of each category for three indexes in Beijing, 2016-2018

Category	WHA_{air}	AQI	AQHI
1	439(2.5%)	2613(14.9%)	329(1.9%)
2	6645(37.9%)	6927(39.5%)	10369(59.1%)
3	5955(34.0%)	4297(24.5%)	6405(36.5%)
4	2780(15.9%)	2121(12.1%)	433(2.5%)
5	1525(8.7%)	1578(9.0%)	/

Table S5. Summary statistics of environmental exposure in Tianjin and Chongqing

		Mean	SD	Min	Q25	Q75	Max
Tianjin	PM_{25} (μg/m^3)	62.6	49.5	1.5	30.5	78.3	508.3
	PM_{10} (μg/m^3)	99.9	64.3	1.5	58.6	123.9	956.2
	NO_2 (μg/m^3)	45.5	22.3	0.4	29.1	58.3	199.0
	SO_2 (μg/m^3)	19.5	23.2	1.4	7.3	21.3	322.6
	O_3 (μg/m^3)	65.8	46.1	1.1	28.8	93.8	350.0
	Relative humidity (%)	55.4	18.6	11.0	40.6	70.5	95.6
	Daily temperature (℃)	13.6	11.4	-15.1	2.1	24.2	33.3
Chongqing	PM_{25} (μg/m^3)	34.1	23.0	2.0	18.2	43.0	218.0
	PM_{10} (μg/m^3)	52.4	30.8	3.8	30.0	66.8	283.3
	NO_2 (μg/m^3)	33.1	15.7	2.6	21.1	42.7	141.4
	SO_2 (μg/m^3)	8.7	4.0	1.4	6.0	10.4	113.3
	O_3 (μg/m^3)	79.7	49.7	1.1	40.5	111.4	358.4
	Relative humidity (%)	78.2	12.6	25.5	7[illegible]	88	100
	Daily temperature (℃)	18.14	8.0	-5.2	11.2	24.6	38.9

Table S6. Comparison of Predictive Performance between WHA_{air}-LSTM and Baseline Models in Beijing

Model	R^2	RMSE	MAE
WHA_{air}-LSTM	0.891	109.24	49.15
GAM	0.7103	126.85	58.88
ARIMA	0.1274	558.61	258.22

Method S1 Exposure-response relationships

The daily counts of total hospital admissions in Beijing from Jan 1st 2013 to Dec 31st 2018 were obtained from the Beijing Municipal Health Commission Information Center, which includes medical records from all the hospitals of Grade II and above in Beijing. We further classified the population into children aged ≤15 years and the elderly aged ≥65 years for the creation of WHA_{air}, to integrate subpopulations vulnerable to air pollution.

We collected hourly concentrations of PM_{10}, $PM_{2.5}$, O_3, SO_2, and NO_2 during the same period from the national air quality monitoring sites within Beijing, and then calculated 24-hour average concentration for PM_{10}, $PM_{2.5}$, SO_2, NO_2, and maximum 8-hour moving average concentration for O_3. Data on temperature, relative humidity and total precipitation were obtained from the ERA5-Land reanalysis dataset, with a spatial resolution of 0.25°×0.25°. We calculated city-level daily mean values by using the area-weighted average of the gridded data.

To obtain the exposure-response relationships between air pollutants and daily hospital admissions, we applied a conditional Poisson regression model within a quasi-Poisson family, which allows better adjustments for autocorrelation and overdispersion [1]. In the main model, the daily total hospital admissions in each district of Beijing were the dependent variable; the daily average concentration of each air pollutant on lag0 to lag7 in each district was the independent variable of interest; we also controlled daily temperature and relative humidity by using natural spline functions (ns) with 4 degrees of freedom; we offset the district-level population. This approach has been demonstrated to be suitable to measure acute health effects associated with short-term exposure to air pollution. We performed subgroup analysis for children and the elderly.

Method S2 Creation of five-category index (WHA_{air})

we created a comprehensive index based on several main principles: first, both the general population and vulnerable subpopulations should be considered in categorizing the index; second, each category of the index could be rationally associated with increasing disease risk; third, distribution of each category of the index should not be totally contradictory with the current national air quality index (AQI) to avoid public misleading. Therefore, we adopted the method introduced in Wong's study[2] and adjusted the cut-off points according to the local feature of excess risk related to air pollution. Briefly speaking, we first calculated a ratio between the median value of excess risk for the vulnerable subpopulations (we chose the one with the greater value between medians of ER_{ch} and ER_{el}) and the median value of ER_{total}. Then, $ER_{adjusted}$ was calculated by dividing ER_{WHO} by the above ratio, which can indicate the health risk of vulnerable populations within the total risk of all population. We classified ER_{total} into five levels: low-level ($\leq 0.5*ER_{adjusted}$), median-level ($> 0.5*ER_{adjusted}$ but $\leq ER_{adjusted}$), high ($> ER_{adjusted}$ but $\leq 1.25*ER_{WHO}$), very high ($>1.25*ER_{WHO}$ but $\leq 1.6*ER_{WHO}$), and serious level ($> 1.6*ER_{WHO}$). The specific cut-offs were identified based on sensitivity analysis to best fit the distribution of AQI. Specifically, the division of thresholds strictly follows the principle of gradually differentiating health risk levels and maximizing the health risk gradient across levels. Moreover, to ensure the robustness of the threshold selection, we calculated the threshold values through an iterative optimization process that distinguishes adjacent risk levels and maximizes the health effect of each level. We then created WHA_{air} with five levels and presented its distribution from 2013 to 2018, alongside AQI and AQHI.

Method S3 LSTM model parameters

Long Short-Term Memory (LSTM) is able to deal with both short-term and long-term dependencies, which is essential to enhance the prediction ability[3]. LSTM, as a variant of Recurrent Neural Network (RNN), was used as a prediction model in this study[4]. The robustness of LSTM was tested by analyzing the daily number of county-level outpatient visits and WHA_{air} features, especially the spatio-temporal attributes. In our main model, current day outpatient visits is the dependent variable, the independent variables included the WHA_{air} index, daily average temperature, daily relative humidity, and daily counts of outpatient visits on previous 1 to 3 days[3].

The dataset was divided into three parts, training, validation and test set. The training set was used to learn the basic patterns of the model's parameters (e.g. WHA_{air}, temperature, relative humidity, and previous-day hospital outpatient visits). The validation set, which played a role during the training phase, helped in tuning hyperparameters (e.g., learning rate) and prevented overfitting by saving model checkpoints. The test set, completely independent of the training process, was used to evaluate the final performance of the model and select the optimal model. Initially, the dataset was split with 80% used for the combined training and validation sets, while the remaining 20% was reserved for testing. In the implementation of the model, the training and validation sets were further divided using ten-fold cross-validation.

The LSTM model processes a 3-day input sequence with daily features including hospital outpatient visits, WHA_{air} and meteorological variables. To identify the optimal predictive model based on the balance between model complexity and predictive performance, we trained three-layer LSTM architectures with progressively scaled hidden units (32, 64, 128, 256, 512) to evaluate the impact of unit expansion. The number of units determines the model's ability to capture temporal patterns and nonlinear relationships in sequential data. A higher number of units endows the model with greater representational power to learn complex data dynamics, but it also raises the risk of overfitting and computational costs. The LSTM model consists of three core components[31]. Specifically, the first layer is an LSTM layer designed to capture temporal dependencies in the data. It operates with three-time steps and the tanh activation function for state updates and sigmoid gates to regulate memory retention and flow. The second layer implements dropout regularization with a conservative rate of 0.1 to mitigate overfitting. The third layer consists of a fully connected layer to perform regression and generate final predictions. The model used the Adam optimizer and Mean Squared Error (MSE) as the loss function with an adaptive learning rate. The initial learning rate was set to 0.001 and the minimum learning rate was 0.00001. Early stopping was used to prevent overfitting.

References

1. Armstrong, B. G., Gasparrini, A. & Tobias, A. Conditional Poisson models: a flexible alternative to conditional logistic case cross-over analysis. *BMC Medical Research Methodology* **14**, 122 (2014).
2. Wong, T. W. *et al.* Developing a risk-based air quality health index. *Atmospheric Environment* **76**, 52–58 (2013).
3. Khatibi, T. & Karampour, N. Predicting the number of hospital admissions due to mental disorders from air pollutants and weather condition descriptors using stacked ensemble of Deep Convolutional models and LSTM models (SEDCMLM). *Journal of Cleaner Production* **280**, 124410 (2021).
4. Van Houdt, G., Mosquera, C. & Nápoles, G. A review on the long short-term memory model. *Artif Intell Rev* **53**, 5929–5955 (2020).