BMC Public Health

RESEARCH

Open Access

Spatial analysis of air pollutant exposure and its association with metabolic diseases using machine learning

Jingjing Liu[1†], Chang Liu[1†], Zhangdaihong Liu[2], Yibin Zhou[3], Xiaoguang Li[1*] and Yang Yang[1*]

Abstract

Background Metabolic diseases (MDs), exemplified by diabetes, hypertension, and dyslipidemia, have become increasingly prevalent with rising living standards, posing significant public health challenges. The MDs are influenced by a complex interplay of genetic factors, lifestyle choices, and socioeconomic conditions. Additionally, environmental pollutants, particularly air pollutants (APs), have attracted increasing attention for their potential role in exacerbating these MDs. However, the impact of APs on the MDs remains unclear. This study introduces a novel machine learning (ML) pipeline, an Algorithm for Spatial Relationships Analysis between Exposome and Metabolic Diseases (ASEMD), to analyze spatial associations between APs and MDs at the prefecture-level city scale in China.

Methods The ASEMD pipeline comprises three main steps: (i) Spatial autocorrelation between APs and MDs is evaluated using Moran's *I* statistic and Local Indicators of Spatial Association (LISA) maps. (ii) dimensionality reduction and spatial similarities identification between APs and MDs clusters using Principal Component Analysis (PCA), k-means clustering, and Jaccard index calculations, further validated through spatial maps. (iii) AP exposure is adjusted by demographic and lifestyle confounders to predict MDs using machine learning models (e.g., eXtreme Gradient Boosting (XGBoost), Random Forest (RF), Decision Tree (DT), LightGBM, and Multi-Layer Perceptron (MLP)). SHAP values are employed to identify key adjusted APs that are linked to MDs. Model performance is evaluated through 10-fold cross-validation using five different metrics. The data utilized include CHARLS (2015) and meteorological data (2013-2015).

Results Significant spatial correlations were found between APs and the prevalence of diabetes, dyslipidemia, and hypertension, with higher prevalence rates observed in alignment with elevated APs concentrations. By adjusting for demographic and lifestyle confounders, APs effectively predicted the risk of developing MDs (AUROC=0.890, 0.877, 0.710 for diabetes, dyslipidemia, and hypertension, respectively). The results showed that CO, $PM_{2.5}$, and AQI were strongly correlated with diabetes, whereas NO_2, $PM_{2.5}$, and PM_{10} were significantly associated with dyslipidemia. For hypertension, CO, O_3, and AQI were mostly correlated. Sensitivity analyses across different regions and different types of APs underscored the robustness of our conclusions.

[†]Jingjing Liu and Chang Liu contributed equally to this work.

*Correspondence:
Xiaoguang Li
lixg@shsmu.edu.cn
Yang Yang
emma002@sjtu.edu.cn
Full list of author information is available at the end of the article

Conclusion The ASEMD pipeline successfully integrates ML models, epidemiological methods, and spatial analysis techniques, providing a robust framework for understanding the complex interactions between APs and MDs. We also identified specific APs, including PM_{10}, CO, and SO_2, as being strongly linked to higher rates of diabetes, dyslipidemia, and hypertension in central and northern cities. Future region-specific public health strategies or interventions, especially in those areas with high pollutant levels, are needed to mitigate air pollution's impact on metabolic health.

Keywords Air Pollutant, Metabolic Disease, Machine Learning

Introduction

According to recent consensus, a wide range of diseases, like hypertension, dyslipidemia, diabetes, non-alcoholic fatty liver disease, osteoporosis, and stroke are now classified as metabolic diseases (MDs) [1, 2]. These MDs are influenced by factors such as unhealthy lifestyles, eating habits and central obesity, which lead to the disturbance of metabolic processes and increase the risk of cardiovascular disease [3]. With rising living standards, the prevalence of MDs has significantly increased, posing a major threat to public health [4]. MDs are influenced by a wide range of factors, spanning molecular to systemic levels. At the molecular level, MDs are regulated by various metabolic enzymes, notably RNA N6-methyladenosine, which plays a critical role in cellular processes [5]. At the system level, MDs are intricately linked with comorbidities and socio-health factors, illustrating how both biological mechanisms and social health determinants collectively impact these conditions [6–9]. Nevertheless, these factors alone may not fully explain the high incidence of MDs.

Environmental pollutants, such as particulate matter, have also been implicated in the etiology of MDs, drawing substantial attention [10, 11]. Existing studies have established a correlation between MDs and environmental pollutants, particularly air pollutants (APs). A systematic review by Chen et al. [12] reported a significant correlation between black carbon, zinc, NO_3^-, SO_4^{2-}, and cardiovascular mortality, with prolonged black carbon exposure notably increasing the prevalence of several MDs. For instance, asthma prevalence increased by 1.215 times, type 2 diabetes by 1.243 times, stroke by 1.141 times. Zhang et al. [13] found that for every 10.0 $\mu g/m^3$ increase in long-term ozone exposure, the prevalence of insulin resistance increased by 1.084 times. Additionally, O_3 and $PM_{2.5}$ exhibited a significant additive interaction on the prevalence of insulin resistance, suggesting that reducing exposure to these pollutants could alleviate the health and economic burden associated with chronic obstructive pulmonary disease. Furthermore, Niedermayer et al. [14] identified significant sex differences in the impact of environmental exposure on MDs, with nitrogen oxides exposure increasing diabetes prevalence by 1.49 times in males but not in females. Despite accumulating evidence linking APs to MDs, the spatial distribution of this relationship and its connection to regional health disparities remain under-explored.

However, the prevalence of MDs is intricately linked to a multitude of factors, including environmental exposures, individual lifestyle choices, genetic predispositions, and socioeconomic conditions [15, 16]. The multifactorial nature of these interactions presents significant challenges for traditional statistical methodologies to elucidate the potential relevance [17, 18]. In recent years, machine learning (ML) techniques have been extensively employed to investigate the complex associations between environmental exposures and MDs, achieving notable advancements in model interpretability in the meantime. By leveraging interpretable models such as Random Forests (RF), Decision Trees, and Gradient Boosting Machines, researchers can predict disease outcomes and identify key predictive variables. For example, Yang et al. [19] utilized Naive Bayes Classifier and Gradient Boosting Machines to analyze a comprehensive dataset encompassing environmental and socioeconomic factors. Their study elucidated the trends in cadmium exposure from 1980 to 2040 and forecasted its future impact on MDs like osteoporosis, diabetes, and cancers. The RF regression model exhibited robust performance (accuracy = 92.1%) and identified critical predictors of cadmium exposure, including per capita rice consumption, zinc production, coal consumption, PM_{10}, lead concentrate production, and urine sample type. This approach provides a valuable and cost-effective method for assessing cadmium exposure in populations, particularly in East Asian countries. Wijaya et al. [20] integrated metagenomic data with seven ML models, including Logistic Regression, Support Vector Machine Linear, and Support Vector Machine Radial Basis Function, to diagnose and predict petroleum-contaminated groundwater. Metagenomics reinforced the predictions and interpretability of the ML framework, which shows great promise as a science-based strategy for on-site monitoring and remediation of environmental pollution. These advanced methodologies provided a more precise scientific basis for understanding the intricate relationships between MDs and environmental exposures, and offered critical insights for future disease prevention strategies and

public health policy development. However, no studies have used ML methods to explore the spatial relationship between APs and MDs.

Therefore, this study proposes a novel ML pipeline, termed ASEMD (Algorithm for Spatial relationships analysis between Exposome and Metabolic Diseases), to explore the spatial correlations between specific APs and MDs. ASEMD addresses the gaps identified in existing research by integrating ML models, epidemiological statistical methods, and spatial geographical mapping techniques. The pipeline incorporates methods such as Uniform Manifold Approximation and Projection (UMAP), K-means clustering, Moran's *I*, Local Indicators of Spatial Association (LISA), linear regression (LR), and Extreme Gradient Boosting (XGBoost). Given that diabetes, dyslipidemia, and hypertension are typical conditions among MDs with high prevalence, we used these diseases as examples to explore the spatial correlations between APs and MDs. The key contributions of this work are as follows:

1. ASEMD highlighted the significant regional impact of APs on MDs, providing new insights into the environmental exposures associated with MDs, and supporting the development of region-specific health strategies and interventions.
2. We demonstrated that *CO*, $PM_{2.5}$, and NO_2 are strongly associated with the prevalence of diabetes, dyslipidemia, and hypertension, and the association between APs and MDs varies spatially across geographic regions.
3. Our study utilized interpretable ML models to identify critical predictors of MDs, improving the accuracy and transparency of environmental health research.

Through ASEMD, we can analyze the relationship between APs and MDs on a spatial scale and examine the interaction between APs and MDs across different regions. The proposed ML framework demonstrates great promise to be used as a science-based strategy for preventing and managing MDs.

Materials and methods

Study population

We utilized data from the China Health and Retirement Longitudinal Study (CHARLS), a large-scale national cohort study that collects a high quality nationally representative sample of Chinese residents aged 45 and older across multiple provinces in China. CHARLS is a longitudinal dynamic cohort study, with the national baseline survey conducted in 2011. Follow-up surveys were conducted in 2011, 2013, 2015, and 2018 across 150 counties and 450 communities (villages) in 28 provinces (including autonomous regions and municipalities) in China [21]. As CHARLS is a dynamic cohort, a few new respondents were recruited in subsequent follow-ups. A more detailed description of the CHARLS dataset has been previously published in [21, 22].

To align the available data of CHARLS and APs data from major Chinese cities, [23], we utilized the 2015 data from the CHARLS cohort. This choice was also made to ensure that the diagnosis of health outcomes occurred after exposure to pollution, thereby reducing the bias introduced by reverse causation to some extent. We included a total of 19,892 participants who had physical examination information. Participants were excluded if they were missing residential location data (at the prefecture level), lived in areas without APs data, or lacked information on blood lipids, blood pressure, or fasting glucose. Finally, 19,973 participants were included in this study. This sample size was determined through power analysis, ensuring sufficient statistical power to detect meaningful effects while minimizing the risk of Type II errors. Furthermore, based on influential demographic factors identified in the literature, we considered 19 confounding factors, including age, sex, BMI, education level, marital status, place of residence, mobility status, smoking, alcohol consumption, and physical activity [24]. According to the CHARLS study, all participants provided informed consent, and the study received approval from the Institutional Review Board of Peking University (IRB00001052-11015).

Diagnosis of diabetes, dyslipidemia, and hypertension

Diagnosis of diabetes

The diagnosis of diabetes was based on the following criteria [25]: fasting plasma glucose $\geq$ 7.0 mmol/L, random plasma glucose $\geq$ 11.1 mmol/L, HbA1c $\geq$ 6.5%, self-reported diabetes diagnosed by a physician, and the use of glucose-lowering drugs/insulin treatment. A diagnosis of diabetes was made if any one of these criteria was met. Specifically, we utilized the 2015 blood test data from the CHARLS cohort to obtain values for fasting plasma glucose, random plasma glucose, and HbA1c. Additionally, we incorporated self-reported diabetes history and treatment status from supplementary questionnaire data to determine the presence of diabetes.

Diagnosis of dyslipidemia

The diagnosis of dyslipidemia was based on the self-reported history of physician-diagnosed dyslipidemia and medication use, with a response of "1. Yes" indicating the presence of dyslipidemia [26].

Diagnosis of hypertension

The diagnosis of hypertension was based on the following criteria: self-reported history of physician-diagnosed hypertension in the questionnaire, with a response of "1. Yes" indicating hypertension; blood pressure measurements taken during the CHARLS survey, with the average of three readings taken at intervals of at least 45 seconds using a digital sphygmomanometer (Omron TM HEM-7200 Monitor, Co., LTD., Dalian, China) as the diagnostic basis. Diagnoses were made using both the old standard from the Chinese Hypertension Guidelines (SBP $\geq$ 140 mmHg and/or DBP $\geq$ 90 mmHg) and the new 2023 standard (SBP $\geq$ 130 mmHg and/or DBP $\geq$ 80 mmHg); self-reported use of modern medical treatments for hypertension and/or medication use to control hypertension in the CHARLS questionnaire, with a response of "1. Yes" to either question indicating hypertension. A diagnosis of hypertension was made if any one of these criteria was met [25].

Measurement of air pollutants

In this study, we collected the monthly average concentrations of six APs ($PM_{2.5}$, PM_{10}, NO_2, SO_2, O_3, CO) from 367 major prefecture-level cities in mainland China (including municipalities and some autonomous prefectures) from 2013 to 2015, sourced from State Meteorological Administration. Additional APs metrics included the 24-hour average concentration for each pollutant, the 8-hour average concentration for O_3, and the Air Quality Index (AQI), resulting in a total of 14 APs metrics. The AQI was calculated according to the "Technical Regulation on Ambient Air Quality Index (Trial)" (HJ 633–2012), which takes into account the aforementioned six common pollutants [27].

To ensure the comprehensive integration of environmental pollution information from each region into our models and analyses, we performed feature extraction. Specifically, for each region and each metric, we calculated the minimum, maximum, mean, standard deviation, and quartiles (25^{th}, 50^{th}, and 75^{th} percentiles) over the three-year period. Combining these, we finally derived 98 APs feature values for each region, which were used as input data for our models.

Algorithm for spatial relationships analysis between exposome and metabolic diseases (ASEMD)

The ASEMD pipeline aims to uncover geographical associations between environmental factors (e.g. pollutants) and diseases, identifying specific pollutants closely linked to MDs prevalence at the regional level. Inputs to the ASEMD pipeline are two geographically aligned datasets: pollution indicators and disease prevalence, provided at various administrative levels (e.g., provincial or prefectural). The details of ASEMD pipeline are illustrated in Fig. 1.

Initially, the MDs prevalence and 19 individual confounders are extracted from the CHARLS cohort. The prevalence of MDs will be calculated at the city level and adjusted by age and gender to obtain the adjusted disease prevalence. Next, the adjusted disease prevalence and APs data will be subjected to spatial autocorrelation analysis using Moran's I index and LISA maps, providing a foundation for validating the spatial relationship between APs and MDs.

$$I = \frac{N}{W} \cdot \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

- N is sample size;
- W is the sum of the spatial weight matrix;
- w_{ij} is the spatial weight between the regions i and j;
- x_i and x_j are the variable values for the regions i and j, respectively;
- $\bar{x}$ is the mean of the variable.

Moran's I measures global spatial autocorrelation, where positive values indicate clustering and negative values suggest dispersion, with higher absolute values representing stronger associations [28]. LISA maps extend this analysis locally, identifying regions with patterns such as high-high (HH), Low-low (LL), high-low (HL) or low-high (LH), providing a more granular view of spatial aggregation. HH and LL mean positive spatial autocorrelation, while HL and LH mean negative spatial autocorrelation. Both are used to evaluate whether a certain variable has a significant spatial aggregation phenomenon in spatial data [29].

Following this, AP data undergo standardization and dimensionality reduction using Principal Component Analysis (PCA). Cities are then clustered based on AP indicators using k-means clustering, yielding k clusters. Similarly, city-level disease prevalence rates are used to classify cities into n clusters across various thresholds. By adjusting both k and n, we form two clustering sets (one from APs data and one from MDs prevalence). The Jaccard index, calculated as

$$Jaccard(X, Y) = \frac{|X \cap Y|}{|X \cup Y|}$$

, is then used to compare these clusters, with the highest Jaccard value indicating optimal alignment between APs and MDs based clustering. In addition, the obtained MDs prevalence and city cluster labels are represented on the maps.

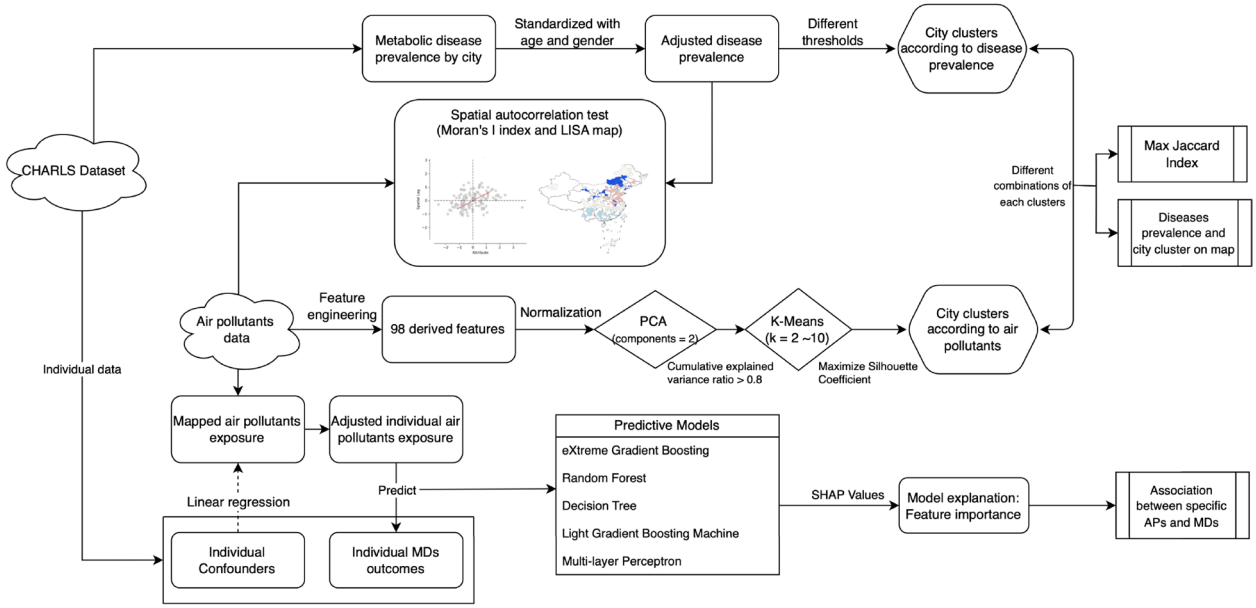


Fig. 1 ASEMD pipeline flowchart. The ASEMD pipeline consists of two main datasets: CHARLS and Air Pollutants data. Firstly, the correlation between APs and MDs is validated from both numerical and visual perspectives using these datasets through Jaccard index and maps. Moran's *I* index and LISA map are applied to examine the spatial autocorrelation of APs and MDs, providing a foundation for understanding their relationship. Secondly, to identify specific APs associated with MDs, linear regression is performed to obtain adjusted individual air pollutants exposure, which is then fed into five predictive models. Thirdly, SHAP values are used to identify key APs and interpret the model outcomes

To investigate the impact of APs on MDs at the individual level, we used a linear regression (LR) model, with various demographic and lifestyle variables as predictors of individual exposure to aindividual-level air pollution. Using these adjusted values, five classification models, eXtreme Gradient Boosting (XGBoost), Random Forest (RF), Decision Tree (DT), Light Gradient Boosting Machine (LightGBM), Multi-Layer Perceptron (MLP), were trained to predict disease prevalence. We applied a 10-fold cross-validation with a data split of 7:2:1 for training, validation, and testing. Model interpretability was provided by Shapley values, which highlighted the significant associations between the APs of interest and MDs.

In this study, data preprocessed through the ASEMD pipeline, along with spatially relevant disease data from the CHARLS cohort, allow for analyzing the association between APs and three MDs at both prefecture and individual scales. Disease diagnosis criteria are based on the CHARLS 2015 data, with prevalence rates age- and sex-standardized across cities to minimize demographic bias. For each disease, prevalence clusters are divided into high and low groups, with cluster thresholds tested at eight levels (0.6–0.95) and environmental clusters adjusted from 3 to 7. The optimal combination of prevalence threshold and environmental cluster count was determined based on the highest Jaccard index. Individual-level analysis then incorporated covariates such as age, gender, BMI, marital status, residence type (urban/rural), exercise, alcohol use, and smoking (see Supplementary Table 1). Five classification models were trained and evaluated through a 10-fold cross-validation with a 7:2:1 split for training, validation, and test sets, and model performance was assessed using on accuracy, AUROC, sensitivity, specificity, precision, and F1 score.

Statistical analysis

For handling missing values in confounders, statistical imputation methods were applied. In general, two strategies were employed to address missing data. First, individuals with missing values for any of the considered variables were excluded from the analysis. Second, when values for specific variables were missing, appropriate imputation methods were applied. For continuous variables, the data were first stratified by gender. For those continuous variables that followed a normal distribution within each group, missing values were imputed using the mean, while for those variables not following a normal distribution, the median was used. For categorical variables, missing values were imputed randomly [30]. Subsequently, continuous variables were standardized using z-scores, and categorical variables were encoded using one-hot encoding [31]. The APs data were calculated as the annual mean values, and there were no

missing values for the annual mean of the cities considered in the study.

The *P*-value derived from Fisher's exact test (calculated using the stats package) was reported alongside each Jaccard index value, with $P < 0.05$ indicating a significant difference between groups. Local Indicators of Spatial Association (LISA) and the spatial autocorrelation statistic Moran's *I* were also calculated, providing insights into spatial clustering and autocorrelation, respectively. These calculations were performed using ArcGIS software (version number). The training and evaluation of the LR and five models were carried out using the xgboost package, lightgbm package and scikit-learn package, respectively. Data sorting and other statistical analysis processes were completed using Python 3.10.

Results

Summarised information about APs and MDs

Table 1 enumerates the APs and the prevalence rates of three MDs, diabetes, hypertension, and dyslipidemia, in 367 cities across the nation. It appears that PM_{10} pollution is significantly more severe on a national scale, whereas *CO* pollution is relatively minimal. Over a 24-hour period, the 25th percentile, median, 75th percentile, and maximum values for the average concentrations of PM_{10} exceed those of other pollutants. Among the MDs, hypertension exhibits the highest national prevalence rate, and diabetes the lowest, and hypertension is 2–3 times more common than diabetes. Following age and gender adjustment, there has been a decrease in the national mean, 25th percentile, 50th percentile, and 75th percentile prevalence rates for these diseases (Table 1). Our results are consistent with previous studies, in 2015, the prevalence of hypertension, diabetes and dyslipidemia was 34.38% [32], 10.9% [33] and 14.1% [34] in the Chinese population aged ≥45 years, with the highest prevalence of hypertension. There is also study show that, after adjusting for age and sex, the prevalence of the diseases decreased [35]. The slightly higher prevalence of our results may be due to the fact that we included all older adults over the age of 45 from CHARLS, not just those aged 45.

Spatial autocorrelation of APs and MDs was reflected by Moran's *I* index and LISA map

In order to explore the spatial correlation between APs and MDs, we first verify the autocorrelation of all factors of APs and MDs. This study calculated the Moran's *I* index for both, and created maps using LISA map. The analysis focused on seven major APs, excluding 8-hour

Table 1 Summary Statistics for air pollutants and metabolic diseases

	Mean	Std	Min	P25	P50	P75	Max	95%CI
APs ($\mu g/m^3$)								
PM2.5	52.828	30.429	0.000	32.240	45.970	65.940	270.850	(52.093, 53.562)
PM2.5 24h	52.685	30.290	0.000	32.060	45.800	65.870	270.030	(51.954, 53.416)
PM10	88.430	47.498	0.000	55.380	77.260	111.280	516.200	(87.284, 89.577)
PM10 24h	88.271	47.491	0.000	55.450	77.260	111.400	516.360	(87.124, 89.417)
NO2	32.648	15.330	0.000	21.230	30.380	41.463	126.680	(32.278, 33.018)
NO2 24h	32.447	15.251	0.000	21.110	30.150	41.240	126.680	(32.079, 32.815)
SO2	28.091	24.171	0.000	13.170	21.110	34.000	255.010	(27.508, 28.675)
SO2 24h	28.076	24.403	0.000	13.170	21.230	33.930	379.500	(27.487, 28.665)
O3	54.751	23.865	0.000	36.000	53.450	70.600	182.080	(54.175, 55.327)
O3 24h	55.101	24.072	0.000	36.330	53.610	71.100	180.780	(54.520, 55.682)
CO	1.137	0.806	0.000	0.733	0.970	1.317	29.000	(1.118, 1.157)
CO 24h	1.122	0.724	0.000	0.726	0.963	1.304	22.000	(1.105, 1.140)
O3_8h	55.405	24.384	0.000	36.190	53.950	71.430	176.250	(54.799, 56.012)
AQI	79.463	36.517	0.000	55.000	71.000	96.000	327.000	(78.582, 80.344)
MDs prevalence(%)								
Diabetes	0.161	0.054	0.054	0.121	0.158	0.196	0.358	(0.152, 0.171)
Diabetes (adjusted)	0.151	0.054	0.048	0.115	0.147	0.184	0.358	(0.142, 0.161)
Hypertension	0.412	0.083	0.190	0.352	0.407	0.459	0.649	(0.397, 0.426)
Hypertension (adjusted)	0.382	0.085	0.168	0.319	0.378	0.439	0.690	(0.367, 0.396)
Dyslipidemia	0.197	0.106	0.020	0.112	0.176	0.256	0.521	(0.178, 0.216)
Dyslipidemia (adjusted)	0.182	0.103	0.026	0.096	0.165	0.246	0.494	(0.163, 0.199)

APs air pollutants, *MDs* metabolic diseases

and 24-hour average values. The results showed that all APs and MDs had significant autocorrelation. In particular, the spatial autocorrelation of PM_{10} (Moran's I index=0.248) in APs is the most significant, while the spatial autocorrelation of dyslipidemia (Moran's I index=0.122) in MDs is the most prominent (Table 2).

The LISA maps of the seven APs and three MDs indicate similar findings (see Supplementary Figure 1 for other APs and MDs). The prevalence rates of diabetes, hypertension, and dyslipidemia in the north-central region exhibit few LL clusters and numerous HL clusters, while the southern regions show few HH clusters and numerous LH clusters. In some western areas, a significant HH clustering of diabetes prevalence is evident. Taking the disease with the highest prevalence, hypertension, and the most severe pollutant, PM_{10}, as examples (Fig. 2), the overall prevalence of hypertension is high in the north-central region, yet it includes a few cities with lower rates that cluster together. Conversely, in the southern regions, although the overall prevalence is lower, there are cities with higher rates clustering together. Overall, the distribution of hypertension rates in the north-central and southern regions shows a clear spatial clustering. Similarly, the distribution of PM_{10} pollution in these regions also exhibits similar spatial clustering. Luo et al. showed that the prevalence of hypertension in the eastern region of China is the highest (32.6%), followed by the northeast region (31.8%), and the lowest is the southwest region (20.1%), that is, compared with the western region, the prevalence of hypertension in the eastern, central and northeast regions is greater, which is the same as our results [36]. Another study showed areas with severe $PM_{2.5}$ pollution in 2015 were mainly concentrated in western Xinjiang, the Beijing-Tianjin-Hebei region, central and eastern Henan Province, and central and western Shandong Province [37]. The results of the distribution of APs are also consistent with the results of our study. Overall, in the north-central region, cities with high prevalence rates of MDs tend to cluster together, while in the southern region, low-prevalence cities are mainly found near high-prevalence cities. In some western areas, cities with a high prevalence of diabetes also form distinct clusters. The spatial distribution of APs such as PM_{10} exhibits similar clustering patterns.

Table 2 Moran's *I* Index and Statistical Analysis for air pollutants and metabolic diseases

Variable	Moran's *I* index	Z value	*P* value
PM2.5	0.216	13.780	0.000
CO	0.126	8.214	0.000
AQI	0.243	15.454	0.000
NO2	0.124	8.090	0.000
O3	0.075	5.103	0.000
PM10	0.248	15.759	0.000
SO2	0.195	12.418	0.000
Diabetes	0.051	3.725	0.000
Dyslipidemia	0.122	8.190	0.000
Hypertension	0.112	7.590	0.000

The *Z*-score is calculated from the Moran's *I* index, the expected index, and the variance, used to test the significance of the Moran's I index. The Expectation Index of all variables is −0.008 and Variance is 0.000. A higher *Z*-score indicates more significant autocorrelation. The *P*-value corresponds to the *Z*-score, with a *P*-value less than 0.05 indicating significant spatial autocorrelation. A *P*-value of 0.000 indicates extremely significant spatial autocorrelation

Spatial correlation of APs and MDs was reflected from two aspects

After confirming that every variables of APs and MDs exhibited spatial autocorrelation, we further explored the spatial correlation between APs and MDs. To determine the optimal number of clusters (m) for geographically related AP indicators, we utilized the Silhouette Coefficient as the decision metric. The analysis revealed that the silhouette score was maximized when $m = 2$ (Supplementary Figure 2). The respective disease prevalence thresholds, maximum Jaccard values, and p-values are presented in Table 3. With $n = 0.6$, the highest Max Jaccard values were observed for Diabetes (0.395), Dyslipidemia (0.377), and Hypertension (0.377), all of which had statistically significant p-values. (0.006, 0.018, 0.018). These findings suggested a strong spatial correlation between these diseases and APs at this threshold. As the threshold increased to $n = 0.65$, the Max Jaccard value for Diabetes decreased to 0.370, but it remained statistically significant ($p = 0.008$), indicating a continued, though slightly diminished, geographical association at this level. Further reductions in Max Jaccard values were observed for Diabetes at higher thresholds: 0.324 at $n = 0.7$ ($p = 0.031$), 0.279 at $n = 0.75$ ($p = 0.039$), and 0.266 at $n = 0.8$ ($p = 0.013$). These results suggested a diminishing spatial relevance as the threshold increases. In summary, the findings underscored a strong geographic correlation between MDs and APs at the lower thresholds ($n = 0.6$ and 0.65), with the spatial association weakening as the threshold rises.

To illustrate the geographical correlation between APs and MDs, both variables were integrated and depicted on a map (Fig. 3). The visualization corroborates the findings presented in the data tables, demonstrating that cities with elevated rates of MDs frequently experience more severe air pollution. The areas most affected by significant air pollution are predominantly located in the northern central parts of the nation and the selected western cities. The prevalence rates of diabetes and dyslipidemia are notably higher in

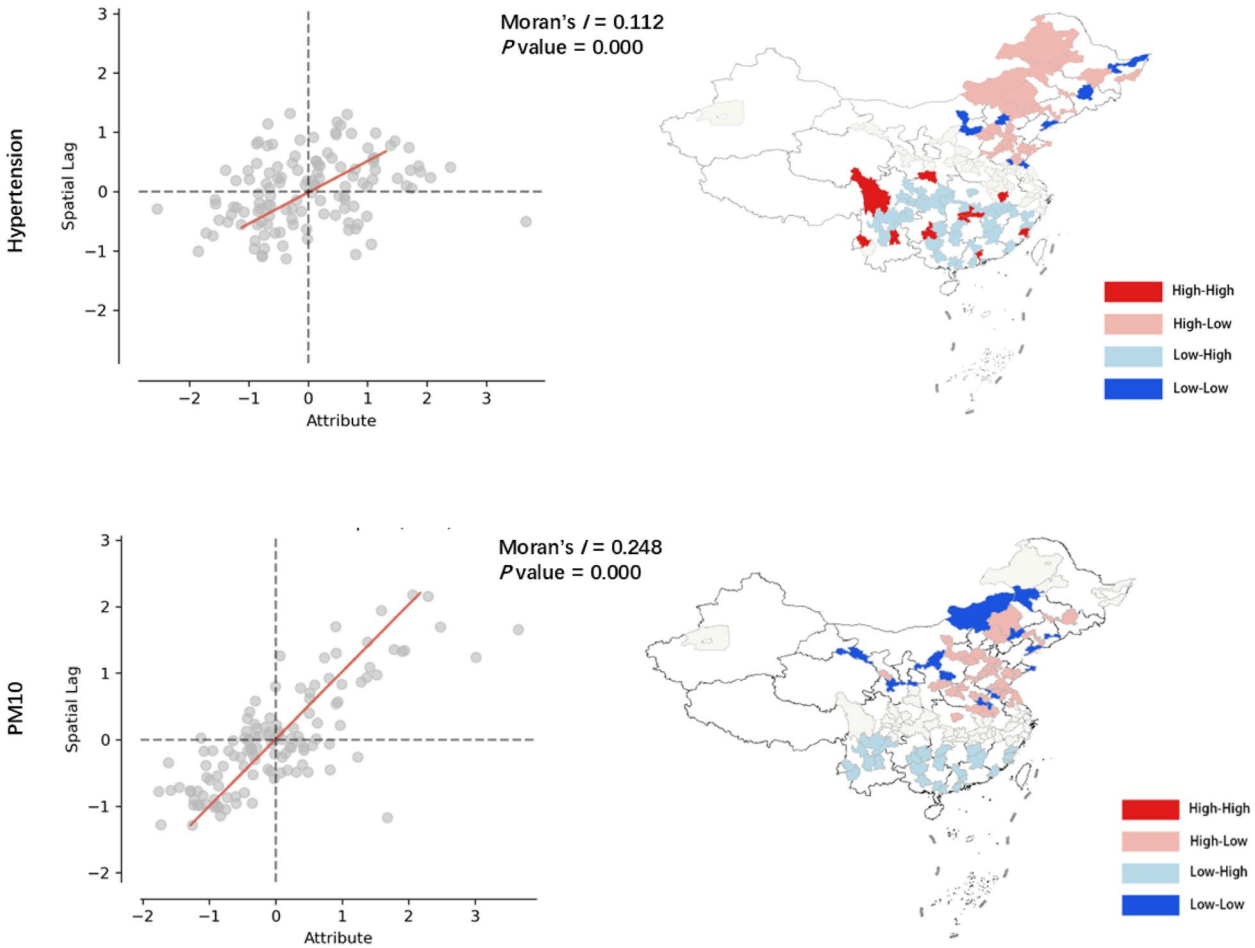


Fig. 2 Spatial analysis benchmarks using Moran's *I* and LISA map for Hypertension and PM_{10}. The left plot is the Moran Scatterplot, with the horizontal axis representing a standardized value for the prevalence of hypertension, and the vertical axis representing the spatial lag value for this urban unit, which is the weighted average of the variable values for other urban units adjacent to this urban unit. The line represents the linear relationship between the spatial lag variable and the prevalence rate of hypertension, and the slope of the line is Moran's *I* value. The four quadrants represent different types of spatial autocorrelation: the upper right quadrant (High-High), lower left quadrant (Low-Low), upper left quadrant (Low-High), and lower right quadrant (High-Low). The right plot is a LISA map, which also illustrates spatial autocorrelations in four categories

Table 3 Analysis of metabolic diseases Incidence by Quantile

Disease	Quantile	Max Jaccard value	*P*-value
Diabetes	quantile_0.6	**0.395**	0.006
Dyslipidemia	quantile_0.6	0.377	0.018
Hypertension	quantile_0.6	0.377	0.018
Diabetes	quantile_0.65	0.370	0.008
Diabetes	quantile_0.7	0.324	0.031
Diabetes	quantile_0.75	0.279	0.039
Diabetes	quantile_0.8	0.266	0.013

Only categories with a *p*-value less than 0.05 from the Fisher test were reported

these central-northern cities, whereas southern cities exhibit lower rates of these conditions. Notably, Jiamusi reports the highest prevalence of diabetes nationally, while Cangzhou has the highest rate of dyslipidemia. Cities across both the northern and southern parts of the country, including Ganzi Tibetan Autonomous Prefecture, Liaocheng, and Yancheng, exhibit high prevalence rates of hypertension. The situation shown on the map is similar to the LISA map results.

Therefore, the spatial correlation between APs and MDs is verified from the two aspects of numerical value (Jaccard Value) and image (geographical map). This provides a basis for the subsequent verification of spatial associations between specific APs and MDs.

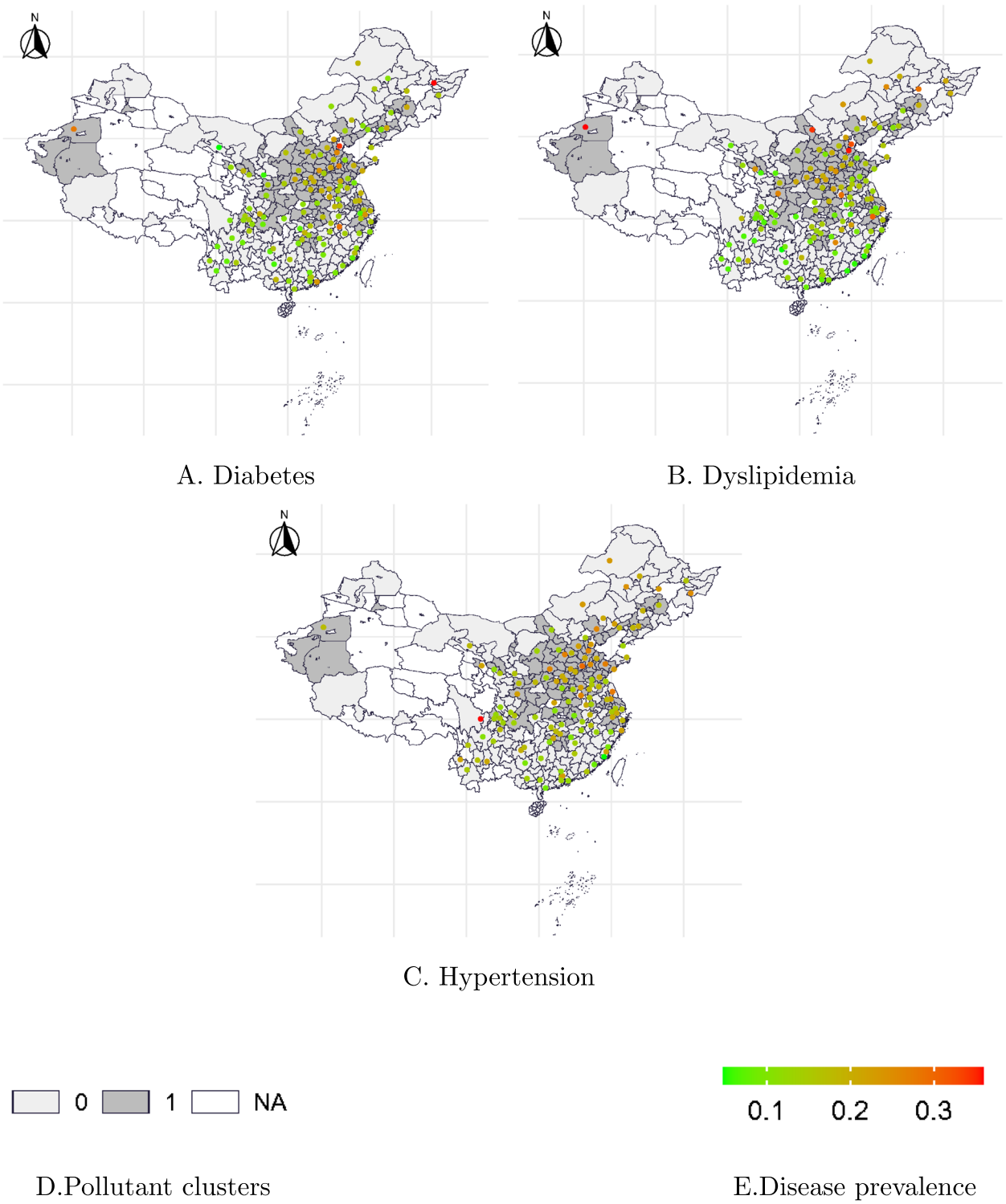


Fig. 3 Prevalence of diabetes, dyslipidemia and hypertension and representation of APs clusters on a map. The black, white and grey blocks on the map represent city clusters based on APs indicators. Colored circles represent MDs Prevalence

Spatial correlation of specific APs and MDs was reflected from ML models

To investigate the specific relationships between APs and MDs, this study first controlled for 19 potential confounding factors, including age, gender, BMI, marital status, place of residence (rural/urban), physical activity, alcohol consumption, and smoking. Second, the adjusted APs data were used as input variables for ML models. Five classification models-XGBoost, RF, DT, LightGBM, MLP-were applied to predict the prevalence of diabetes, dyslipidemia, and hypertension. Third, SHAP values were utilized to identify the top 10 influential APs with predictive effects for each of the three MDs (Fig. 4). The results showed that different APs can effectively predict the prevelence of these MDs (Table 4). Specifically, PM_{10} had the most significant spatial correlation with diabetes (14.7%), *CO* with dyslipidemia (44.3%), and SO_2 with hypertension (22.9%).

Overall, the ML models performed well in predicting the three diseases: diabetes (AUROC = 0.711–0.890), dyslipidemia (AUROC = 0.706–0.877), and hypertension (AUROC = 0.603–0.710). Among them, the XGBoost model performed the best, achieving AUROCs of 0.890 for diabetes, 0.877 for dyslipidemia, and 0.710 for hypertension.

Sensitivity analysis

To validate the stability of our results, we employed the ASEMD method and selected the most effective XGBoost model for our sensitivity analysis. This analysis was stratified into two main categories: urban-rural population stratification and APs stratification. Across these strata, all Max Jaccard values exceeded 0.2, indicating a high level of spatial association (Table 5). In the urban-rural stratification, it was observed that the prevalence of dyslipidemia and hypertension showed a stronger spatial association with APs in urban populations. In contrast, the prevalence of diabetes exhibited a greater correlation with APs in rural populations, with dyslipidemia in urban areas demonstrating the strongest association. The pollutant stratification analysis revealed robust spatial correlations between seven types of pollutants and the three diseases studied. Notably, PM_{10} showed the strongest association with all three diseases, and specifically, the correlation between diabetes and NO_2 was the most pronounced.

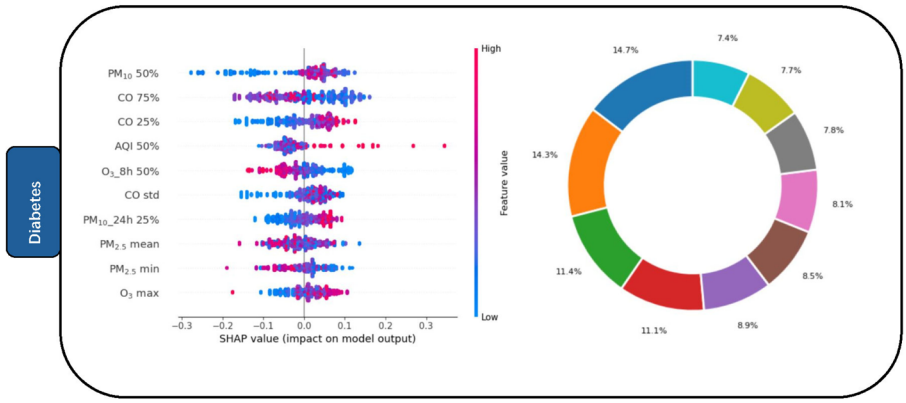
Additionally, the AUROC values under both stratification methods were analyzed for the XGBoost model, which demonstrated excellent performance (Table 6). Particularly, the adjusted values of APs for age and gender accurately diagnosed diabetes in rural populations most effectively. Ozone (O_3) performed best in diagnosing all three diseases, especially diabetes, where it achieved an AUROC of 0.858.

These results indicate that ASEMD maintained robustness under both urban-rural and APs stratification analyses. Furthermore, the XGBoost model showed high diagnostic efficacy in both stratification settings.

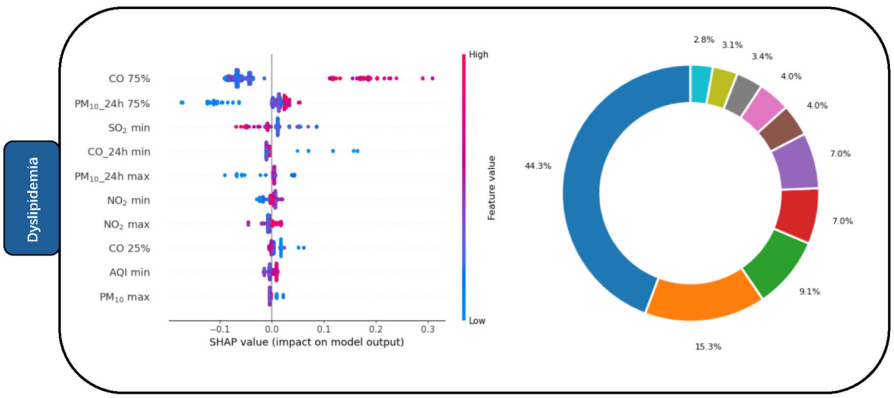
Discussion

In this study, we employed a novel ASEMD pipeline to investigate the spatial associations between common APs and MDs (diabetes, hypertension, and dyslipidemia). ASEMD utilizes high-dimensional APs data to perform PCA, K-means clustering, and Jaccard index calculation to evaluate and rank the similarity between pollutant clusters and the actual geographical distribution of chronic disease prevalence. Additionally, an ML-based prediction model for the effects of pollutants on chronic diseases was constructed, thereby assessing the spatial associations between them in a multi-layered and multidimensional manner. Our findings reveal that there is a spatial correlation between APs and the prevalence of diabetes, dyslipidemia, and hypertension, indicating that regions with relatively higher concentrations of APs also have higher prevalence rates of these MDs. Moreover, our results demonstrate that even after accounting for various demographic and lifestyle confounders, APs can be effectively modelled to predict an individual's risk of MDs while identifying the most significant association between specific AP and MDs. In detail, PM_{10} exhibited the strongest associations with diabetes, while *CO* were the most significant pollutants for dyslipidemia. SO_2 showed relatively strong associations for hypertension. Sensitivity analyses that distinguish between regions and individual pollutants were consistent with the main findings, suggesting a degree of robustness in our conclusions.

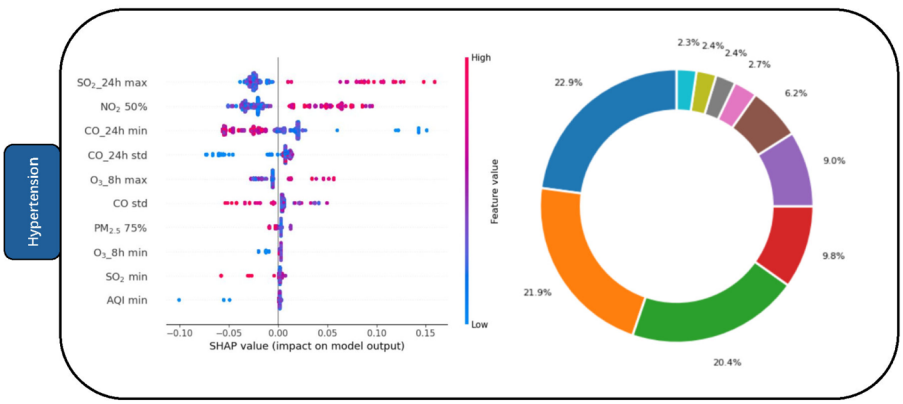
Our primary findings align with previous research indicating that APs are associated with MDs and comorbidities, such as cardiovascular diseases. For instance, particulate matters (including PM_1, $PM_{2.5}$, and PM_{10}) were important pollutants associated with the three diseases considered in our study. Previous studies have shown that they were closely associated with the risk of MDs such as hypertension, coronary heart disease, diabetes, dyslipidemia, and metabolic syndrome [38–42]. PM_{10}, capable of penetrating and lodging deep in the lungs, can cause irritation, inflammation, and damage to the respiratory tract lining, primarily affecting the respiratory system. A toxicological study suggested that PM_{10} may induce cardiovascular toxicity by elevating ROS levels in the body [43]. Particulate matter with a diameter of 2.5 micrometres or smaller (P$M_{2.5}$) is even more harmful, as it can penetrate the lung barrier and enter



A.Diabetes



B.Dyslipidemia



C.Hypertension

Fig. 4 Top 10 features importance of MDs using SHAP. The subplots on the left are Shapley additive explanations, where the pink indicates that the feature has a higher SHAP value; blue indicates feature has a lower SHAP value. The top 10 APs features with the most significant impact on MDs prediction are listed. The higher ranking of an AP feature, the greater effects on the model's prediction. The subplots one the right illustrate the relative importance ratios of these ten APs

the bloodstream, affecting all major organs and thereby increasing the risk of various cardiovascular events and mortality [44, 45]. The primary mechanisms of $PM_{2.5}$'s health damage could involve oxidative stress in the lungs, systemic inflammation, vascular dysfunction, and atherosclerosis. Long-term exposure to $PM_{2.5}$ has been linked

Table 4 Performance metrics for different models predicting metabolic diseases outcomes

Metric	XGBoost	RF	DT	LightGBM	MLP	*P*-value
Diabetes						
Accuracy	**0.814 ± 0.007**	0.659 ± 0.007	0.749 ± 0.008	0.726 ± 0.008	0.751 ± 0.006	< 0.0001
AUROC	**0.890 ± 0.005**	0.711 ± 0.008	0.749 ± 0.008	0.797 ± 0.008	0.835 ± 0.011	< 0.0001
Sensitivity	**0.847 ± 0.009**	0.727 ± 0.017	0.765 ± 0.012	0.783 ± 0.009	0.778 ± 0.067	0.0001
Specificity	**0.781 ± 0.008**	0.591 ± 0.017	0.732 ± 0.012	0.668 ± 0.010	0.725 ± 0.066	0.0009
Precision	**0.794 ± 0.012**	0.640 ± 0.011	0.741 ± 0.008	0.702 ± 0.007	0.742 ± 0.031	< 0.0001
F1 Score	**0.820 ± 0.009**	0.680 ± 0.006	0.753 ± 0.008	0.740 ± 0.008	0.757 ± 0.017	0.0001
Dyslipidemia						
Accuracy	**0.804 ± 0.008**	0.640 ± 0.014	0.735 ± 0.008	0.723 ± 0.009	0.747 ± 0.008	< 0.0001
AUROC	**0.877 ± 0.009**	0.706 ± 0.009	0.736 ± 0.009	0.791 ± 0.010	0.825 ± 0.009	< 0.0001
Sensitivity	**0.836 ± 0.012**	0.594 ± 0.019	0.749 ± 0.011	0.762 ± 0.016	0.788 ± 0.041	0.0003
Specificity	**0.772 ± 0.009**	0.686 ± 0.013	0.721 ± 0.011	0.684 ± 0.020	0.707 ± 0.012	< 0.0001
Precision	**0.786 ± 0.008**	0.654 ± 0.018	0.728 ± 0.009	0.707 ± 0.012	0.730 ± 0.018	< 0.0001
F1 Score	**0.810 ± 0.008**	0.622 ± 0.016	0.738 ± 0.009	0.734 ± 0.009	0.757 ± 0.013	< 0.0001
Hypertension						
Accuracy	**0.655 ± 0.010**	0.620 ± 0.011	0.603 ± 0.009	0.639 ± 0.013	0.638 ± 0.010	< 0.0001
AUROC	**0.710 ± 0.012**	0.665 ± 0.013	0.603 ± 0.009	0.691 ± 0.012	0.677 ± 0.013	< 0.0001
Sensitivity	**0.671 ± 0.015**	0.652 ± 0.011	0.618 ± 0.014	0.665 ± 0.019	0.678 ± 0.029	< 0.0001
Specificity	**0.638 ± 0.012**	0.588 ± 0.018	0.589 ± 0.015	0.612 ± 0.014	0.597 ± 0.023	0.0004
Precision	**0.650 ± 0.010**	0.613 ± 0.015	0.601 ± 0.009	0.632 ± 0.012	0.628 ± 0.009	< 0.0001
F1 Score	**0.660 ± 0.009**	0.631 ± 0.010	0.609 ± 0.010	0.648 ± 0.014	0.652 ± 0.014	< 0.0001

XGBoost is Extreme Gradient Boosting, RF is Random Forest, DT is Decision Tree, LightGBM is Light Gradient Boosting Machine, MLP is Multi-Layer Perceptron. The *P*-values were obtained by the Friedman test

Table 5 Max Jaccard value for sensitivity analysis of rural urban stratification and air pollutants stratification

Disease	Population stratification		APs stratification						
	Urban	**Rural**	$PM_{2.5}$	PM_{10}	***CO***	NO_2	O_3	SO_2	***AQI***
Diabetes	0.377	**0.386**	0.410	0.403	0.273	**0.425**	0.250	0.303	0.405
Dyslipidemia	**0.390**	0.333	0.254	0.384	0.333	0.262	0.293	0.217	0.275
Hypertension	0.381	0.244	0.167	0.309	0.279	0.217	0.222	0.222	0.388

AQI stands for Air Quality Index

Subgroup1 is the rural town stratification and Subgroup2 is the air pollutants stratification

Table 6 XGBoost model AUROC for sensitivity analysis of rural-urban stratification and air pollutants stratification

Disease	Population stratification		APs stratification						
	Urban	**Rural**	$PM_{2.5}$	PM_{10}	***CO***	NO_2	O_3	SO_2	***AQI***
Diabetes	**0.908 ± 0.008**	**0.911 ± 0.005**	0.848 ± 0.007	0.852 ± 0.009	0.841 ± 0.008	0.849 ± 0.008	**0.858 ± 0.008**	0.832 ± 0.006	0.837 ± 0.008
Dyslipidemia	0.869 ± 0.010	0.905 ± 0.007	0.833 ± 0.006	0.838 ± 0.0056	0.825 ± 0.009	0.841 ± 0.008	0.850 ± 0.009	0.823 ± 0.010	0.823 ± 0.007
Hypertension	0.720 ± 0.012	0.719 ± 0.014	0.667 ± 0.016	0.672 ± 0.011	0.672 ± 0.008	0.684 ± 0.012	0.697 ± 0.010	0.678 ± 0.014	0.665 ± 0.013

AQI stands for Air Quality Index

to metabolic syndrome, dyslipidemia, and impaired fasting glucose [46–48]. A review reported that *CO* exposure disrupted lipid metabolism and increased oxidative stress, contributing to dyslipidemia, particularly elevated cholesterol and triglycerides [49]. By binding to hemoglobin, *CO* reduces oxygen transport, creating an anoxic environment that promotes fatty acid oxidation [50]. Additionally, long-term exposure to SO_2 and other pollutants may affect vascular endothelial function through oxidative stress, chronic inflammatory response and other mechanisms [51, 52]. Studies have also demonstrated that exposure to SO_2 can directly impact the cardiovascular system by triggering the sympathetic nervous system, leading to an increase in blood pressure. For individuals already living with hypertension, exposure to SO_2 and other APs can exacerbate their condition, resulting in more pronounced fluctuations in blood pressure [53, 54].

Furthermore, our study underscores the spatial heterogeneity or regional variability in the impact of APs on MDs. For example, while there is a high correlation between pollutant clusters and disease prevalence clusters, the spatial map does not display a uniform pattern of association, indicating that the impact of APs is not evenly distributed across different regions. We hypothesize that this may be due to variations in the spatial patterns of pollutant exposure and other regional factors. For instance, the prevalence of hypertension tends to be higher in northern China, independent of the distribution of APs, which we believe is likely due to factors such as diet and lifestyle [55, 56]. Additionally, socioeconomic factors also contribute to this phenomenon, as previous studies have shown that areas with higher levels of socioeconomic deprivation may suffer greater health impacts from MDs due to limited access to healthcare, higher baseline health risks, and greater exposure to APs [57, 58].

Compared to traditional epidemiological methods, this innovative ASEMD pipeline that we propose has several significant advantages. It allows for a more detailed analysis of spatial association patterns between exposure and disease, and the incorporation of individual-level confounding factors into ML models, thereby mitigating the influence of more confounders and reducing ecological fallacy to some extent [59]. Traditional epidemiological methods often rely on simpler models that may not fully capture the spatial complexity of environmental exposure. They typically focus on relationships at the individual level (e.g., individual behavior, lifestyle) or at the regional level (e.g., air pollution and regional disease prevalence). However, they may overlook spatial correlations. The proposed ASEMD provided a more powerful framework for understanding the complex relationships between environmental factors and health outcomes. This method not only improves the reliability of the predictions but also enhances the interpretability of the results.

Nevertheless, it is important to note that our study has some limitations. First, due to data granularity in CHARLS, we used data aggregated at the prefectural city level. Given the relatively size of prefectural cities in China and the considerable internal heterogeneity, future research should focus on incorporating finer-grained data, such as individual health records and personal exposure assessments, to improve the accuracy of the analysis and increase confidence in our results. Second, our study employed a cross-sectional data sampling method, which limits our ability to draw causal inferences. Furthermore, the covariates included in our study may not fully cover all risk factors for MDs, which could affect the robustness of the associations observed in our research. Finally, there was potential selection bias in the CHARLS cohort. The dataset predominantly consisted of older adults (60 years and older) in China, which limited the generalizability of the findings to younger populations or those with different health conditions. Additionally, since participation was voluntary, it is likely that individuals with better health or higher socioeconomic status are under-presented, potentially distorting the results.

Conclusion

This study presents the ASEMD pipeline as an innovative approach to explore the spatial correlations between APs and MDs. Our findings underscore the significant spatial associations between diabetes, dyslipidemia, hypertension, and various APs, particularly PM_{10} with diabetes, *CO* with dyslipidemia, and SO_2 with hypertension. The ASEMD algorithm successfully integrates ML models, epidemiological methods, and spatial analysis techniques, providing a robust framework to understand the complex interactions between APs and MDs. The results highlight the uneven spatial distribution of these associations, indicating regional variability. The instability of this spatial correlation is likely influenced by other factors such as socioeconomic conditions, lifestyle, and diet. Moreover, the study's use of interpretable ML models offers valuable insights into the key predictors of MDs prevalence, enhancing the transparency and reliability of the findings.

Our findings support the need for targeted, region-specific public health strategies and interventions, especially in areas with high levels of these APs, to mitigate the effects of air pollution on metabolic health. Future research should focus on enhancing the ASEMD pipeline by integrating more fine-grained data, such as individual level pollutant exposures and health records, and

conducting longitudinal studies to better establish causality and improve the accuracy of disease risk predictions based on environmental exposures.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12889-025-22077-9.

Additional analyses can be found in the Supplementary Files.

Materials availability

Not applicable.

Authors' contributions

YY and XGL conceptualized the study and convened the full team of authors. JJL and CL obtained the APs information on the State Meteorological Administration and the metabolic diseases data on CHARLS. JJL, CL and ZDHL collected, processed and analyzed the data according to the pipeline. JJL, CL, ZDHL, YBZ participated in weekly or biweekly meetings coordinated by YY and XGL. ZDHL, YBZ, YY and XGL provided important insights into the conceptual approach. JJL and CL drafted the manuscript. ZDHL, YBZ, YY and XGL critically revised the manuscript. All authors read, edited, critically revised, and approved the final manuscript.

Funding

This work is supported by Shanghai Natural Science Foundation (Grant No. 23ZR1436400).

Data availability

The data supporting the findings of this study are publicly available data sets CHARLS and 2013-2015 National Weather Service Air Pollutants Data.

Code availability

Not applicable.

Declarations

Ethics approval and consent to participate

Each round of CHARLS investigation was approved by the Biomedical Ethics Committee of Peking University. The field work plan of this round of household questionnaire survey has been approved, and the approval number is IRB00001052-11015. All participants provided written informed consent.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]School of Public Health, Shanghai Jiao Tong University School of Medicine, 227 South Chongqing Road, Shanghai 200025, China. [2]Oxford Suzhou Center for Advanced Research, Building A, 388 Ruo Shui Road, Suzhou Industrial Park, Suzhou 215123, China. [3]Shanghai Minhang District Center for Disease Control and Prevention, No.965, Zhongyi Road, Qibao Town, Minhang District, Shanghai 201101, China.

Received: 6 November 2024 Accepted: 24 February 2025
Published online: 01 March 2025

References

1. Wang X, Zhang C, Zhao G, Yang K, Tao L. Obesity and lipid metabolism in the development of osteoporosis. Int J Mol Med. 2024;54(1):1–11.
2. Luciani L, Pedrelli M, Parini P. Modification of lipoprotein metabolism and function driving atherogenesis in diabetes. Atherosclerosis. 2024;394:117545.
3. Qiu L, Wang W, Sa R, Liu F. Prevalence and risk factors of hypertension, diabetes, and dyslipidemia among adults in Northwest China. Int J Hypertens. 2021;2021(1):5528007.
4. Chen X, Zhang L, Chen W. Global, regional, and national burdens of type 1 and type 2 diabetes mellitus in adolescents from 1990 to 2021, with forecasts to 2030: a systematic analysis of the global burden of disease study 2021. BMC Med. 2025;23(1):48.
5. Li Y, Wang J, Huang C, Shen M, Zhan H, Xu K. RNA N 6-methyladenosine: a promising molecular target in metabolic diseases. Cell Biosci. 2020;10(1):19.
6. Zlokovic BV, Gottesman RF, Bernstein KE, Seshadri S, McKee A, Snyder H, et al. Vascular contributions to cognitive impairment and dementia (VCID): A report from the 2018 National Heart, Lung, and Blood Institute and National Institute of Neurological Disorders and Stroke Workshop. Alzheim Dement. 2020;16(12):1714–33.
7. Johansen MC, Ye W, Gross A, Gottesman RF, Han D, Whitney R, et al. Association between acute myocardial infarction and cognition. JAMA Neurol. 2023;80(7):723–31.
8. Yaffe K, Weston AL, Blackwell T, Krueger KA. The metabolic syndrome and development of cognitive impairment among older women. Arch Neurol. 2009;66(3):324–8.
9. Singh-Manoux A, Dugravot A, Brunner E, Kumari M, Shipley M, Elbaz A, et al. Interleukin-6 and C-reactive protein as predictors of cognitive decline in late midlife. Neurology. 2014;83(6):486–93.
10. Zhang X, Li Z, Hu R, Liu X, Yang W, Wu Y, et al. Exposure memory and susceptibility to ambient PM2. 5: A perspective from hepatic cholesterol and bile acid metabolism. Ecotoxicol Environ Saf. 2024;280:116589.
11. Li Y, Lv Y, Jiang Z, Ma C, Li R, Zhao M, et al. Association of co-exposure to organophosphate esters and per-and polyfluoroalkyl substances and mixture with cardiovascular-kidney-liver-metabolic biomarkers among Chinese adults. Ecotoxicol Environ Saf. 2024;280:116524.
12. Chen S, Liu D, Huang L, Guo C, Gao X, Xu Z, et al. Global associations between long-term exposure to PM2. 5 constituents and health: A systematic review and meta-analysis of cohort studies. J Hazard Mater. 2024;474:134715.
13. Zhang Z, Luan C, Wang C, Li T, Wu Y, Huang X, et al. Insulin resistance and its relationship with long-term exposure to ozone: Data based on a national population cohort. J Hazard Mater. 2024;472:134504.
14. Niedermayer F, Wolf K, Zhang S, Dallavalle M, Nikolaou N, Schwettmann L, et al. Sex-specific associations of environmental exposures with prevalent diabetes and obesity-Results from the KORA Fit study. Environ Res. 2024;252:118965.
15. Noubiap JJ, Nansseu JR, Lontchi-Yimagou E, Nkeck JR, Nyaga UF, Ngouo AT, et al. Geographic distribution of metabolic syndrome and its components in the general adult population: A meta-analysis of global data from 28 million individuals. Diabetes Res Clin Pract. 2022;188:109924.
16. Chew NW, Ng CH, Tan DJH, Kong G, Lin C, Chin YH, et al. The global burden of metabolic disease: Data from 2000 to 2019. Cell Metab. 2023;35(3):414–28.
17. Cliff OM, Bryant AG, Lizier JT, Tsuchiya N, Fulcher BD. Unifying pairwise interactions in complex dynamics. Nat Comput Sci. 2023;3(10):883–93.
18. Miao J, Wu Y, Lu Q. Statistical methods for gene-environment interaction analysis. Wiley Interdiscip Rev Comput Stat. 2024;16(1):e1635.
19. Yang Y, Gu Y, Zhang Y, Zhou Q, Zhang S, Wang P, et al. Spatial-temporal mapping of urine cadmium levels in China during 1980–2040: Dietary improvements lower exposure amid rising pollution. J Hazard Mater. 2024;473:134693.
20. Wijaya J, Park J, Yang Y, Siddiqui SI, Oh S. A metagenome-derived artificial intelligence modeling framework advances the predictive diagnosis and interpretation of petroleum-polluted groundwater. J Hazard Mater. 2024;472:134513.
21. Zhao Y, Hu Y, Smith JP, Strauss J, Yang G. Cohort profile: the China health and retirement longitudinal study (CHARLS). Int J Epidemiol. 2014;43(1):61–8.

22. Zhao Y, Strauss J, Yang G, Giles J, Hu P, Hu Y, et al. China health and retirement longitudinal study–2011–2012 national baseline users' guide. 2013;2:1–56. Beijing: National School of Development, Peking University.
23. Chen X, Crimmins E, Hu P, Kim JK, Meng Q, Strauss J, et al. Venous blood-based biomarkers in the China health and retirement longitudinal study: rationale, design, and results from the 2015 wave. Am J Epidemiol. 2019;188(11):1871–7.
24. Fotakis C, Amanatidou AI, Kafyra M, Andreou V, Kalafati IP, Zervou M, et al. Circulatory Metabolite Ratios as Indicators of Lifestyle Risk Factors Based on a Greek NAFLD Case-Control Study. Nutrients. 2024;16(8):1235.
25. Lin L, Wang HH, Lu C, Chen W, Guo VY. Adverse childhood experiences and subsequent chronic diseases among middle-aged or older adults in China and associations with demographic and socioeconomic characteristics. JAMA Netw Open. 2021;4(10):e2130143–e2130143.
26. Wu Y, Yang Y, Zhang J, Liu S, Zhuang W. The change of triglyceride-glucose index may predict incidence of stroke in the general population over 45 years old. Cardiovasc Diabetol. 2023;22(1):132.
27. of Ecology M, of the People's Republic of China E. Technical Regulation on Ambient Air Quality Index (on trial). HJ. 2016:633–2012.
28. Liang Y, Gong Z, Guo J, Cheng Q, Yao Z. Spatiotemporal analysis of the morbidity of global Omicron from November 2021 to February 2022. J Med Virol. 2022;94(11):5354–62.
29. Zhang Y, Sun T, Wang L, Huang B, Pan X, Song W, et al. Portraying on-road CO2 concentrations using street view panoramas and ensemble learning. Sci Total Environ. 2024;946:174326.
30. Afkanpour M, Hosseinzadeh E, Tabesh H. Identify the most appropriate imputation method for handling missing values in clinical structured datasets: a systematic review. BMC Med Res Methodol. 2024;24(1):188.
31. Singh VK, Maurya NS, Mani A, Yadav RS. Machine learning method using position-specific mutation based classification outperforms one hot coding for disease severity prediction in haemophilia 'A'. Genomics. 2020;112(6):5122–8.
32. Zhang H, Zhou XD, Shapiro MD, Lip GY, Tilg H, Valenti L, et al. Global burden of metabolic diseases, 1990–2021. Metabolism. 2024;160:155999.
33. Li Y, Teng D, Shi X, Qin G, Qin Y, Quan H, et al. Prevalence of diabetes recorded in mainland China using 2018 diagnostic criteria from the American Diabetes Association: national cross sectional study. BMJ. 2020;369:m997.
34. Li S, Zhang L, Wang X, Chen Z, Dong Y, Zheng C, et al. Status of dyslipidemia among adults aged 35 years and above in China. Chin Circ J. 2019;34(7):681–7.
35. Chong B, Kong G, Shankar K, Chew HJ, Lin C, Goh R, et al. The global syndemic of metabolic diseases in the young adult population: A consortium of trends and projections from the Global Burden of Disease 2000–2019. Metabolism. 2023;141:155402.
36. Luo Y, Xia F, Yu X, Li P, Huang W, Zhang W. Long-term trends and regional variations of hypertension incidence in China: a prospective cohort study from the China Health and Nutrition Survey, 1991–2015. BMJ Open. 2021;11(1):e042053.
37. Lei J, Zhou H, Lai Z, Bai L, Chen Z. Analysis of spatio-temporal characteristic of PM2. 5 concentrations of Chinese cities: 2015–2017. Acta Sci Circumstantiae. 2018;38:3816–3825.
38. Sun M, Li T, Sun Q, Ren X, Sun Z. Duan J. Associations of long-term particulate matter exposure with cardiometabolic diseases: A systematic review and meta-analysis. Sci Total Environ; 2023. p. 166010.
39. Zhou Q, Li X, Zhang J, Duan Z, Mao S, Wei J, et al. Long-term exposure to PM1 is associated with increased prevalence of metabolic diseases: evidence from a nationwide study in 123 Chinese cities. Environ Sci Pollut Res. 2024;31(1):549–63.
40. Chen C, Wang X, Lv C, Li W, Ma D, Zhang Q, et al. The Effect of Air Pollution on Hospitalization of Individuals with Respiratory and Cardiovascular Diseases in Jinan, China. Medicine. 2019;98(22):e15634.
41. Feng W, Li H, Wang S, Van Halm-Lutterodt N, An J, Liu Y, et al. Short-Term PM10 and Emergency Department Admissions for Selective Cardiovascular and Respiratory Diseases in Beijing. China Sci Total Environ. 2019;657:213–21. https://doi.org/10.1016/j.scitotenv.2018.12.066.
42. Ma J, Zhang J, Zhang Y, Wang Z. Causal effects of noise and air pollution on multiple diseases highlight the dual role of inflammatory factors in ambient exposures. Sci Total Environ. 2024;951:175743.
43. Cen J, Jia ZI, Zhu Cy, Wang Xf, Zhang F, Chen Wy, et al. Particulate Matter (PM10) Induces Cardiovascular Developmental Toxicity in Zebrafish Embryos and Larvae via the ERS, Nrf2 and Wnt Pathways. Chemosphere. 2020;250:126288. https://doi.org/10.1016/j.chemosphere.2020.126288.
44. Alexeeff SE, Liao NS, Liu X, Van Den Eeden SK, Sidney S. Long-Term PM2.5 Exposure and Risks of Ischemic Heart Disease and Stroke Events: Review and Meta-Analysis. J Am Heart Assoc. 2021;10(1):e016890. https://doi.org/10.1161/JAHA.120.016890.
45. Rajagopalan S, Landrigan PJ. Pollution and the Heart. N Engl J Med. 2021;385(20):1881–92. https://doi.org/10.1056/NEJMra2030281.
46. Mazidi M, Speakman JR. Ambient Particulate Air Pollution (PM2.5) Is Associated with the Ratio of Type 2 Diabetes to Obesity. Sci Rep. 2017;7(1):9144. https://doi.org/10.1038/s41598-017-08287-1.
47. Ma R, Zhang Y, Sun Z, Xu D, Li T. Effects of Ambient Particulate Matter on Fasting Blood Glucose: A Systematic Review and Meta-Analysis. Environ Pollut. 2020;258:113589. https://doi.org/10.1016/j.envpol.2019.113589.
48. Zheng Xy, Tang Sl, Liu T, Wang Y, Xu Xj, Xiao N, et al. Effects of Long-Term PM2.5 Exposure on Metabolic Syndrome among Adults and Elderly in Guangdong, China. Environ Health. 2022;21(1):84. https://doi.org/10.1186/s12940-022-00888-2.
49. Wang C, Meng Xc, Huang C, Wang J, Liao Yh, Huang Y, et al. Association between ambient air pollutants and lipid profile: A systematic review and meta-analysis. Ecotoxicol Environ Saf. 2023;262:115140.
50. Ren QW, Teng THK, Ouwerkerk W, Tse YK, Tsang CTW, Wu MZ, et al. Triglyceride levels and its association with all-cause mortality and cardiovascular outcomes among patients with heart failure. Nat Commun. 2025;16(1):1–11.
51. Liu X, Zhou H, Zhang H, Jin H, He Y. Advances in the research of sulfur dioxide and pulmonary hypertension. Front Pharmacol. 2023;14:1282403.
52. Hudson J, Farkas L. Epigenetic regulation of endothelial dysfunction and inflammation in pulmonary arterial hypertension. Int J Mol Sci. 2021;22(22):12098.
53. Ulusoy Ş, Özkan G, Varol G, Erdem Y, Derici Ü, Yılmaz R, et al. The Effect of Ambient Air Pollution on Office, Home, and 24-Hour Ambulatory Blood Pressure Measurements. Am J Hypertens. 2023;36(8):431–8.
54. Meo SA, Shaikh N, Alotaibi M. Association between air pollutants particulate matter (PM2. 5, PM10), nitrogen dioxide (NO2), sulfur dioxide (SO2), volatile organic compounds (VOCs), ground-level ozone (O3) and hypertension. J King Saud Univ-Sci. 2024;36(11):103531.
55. Fling C, De Marco T, Kime NA, Lammi MR, Oppegard LJ, Ryan JJ, et al. Regional Variation in Pulmonary Arterial Hypertension in the United States: The Pulmonary Hypertension Association Registry. Ann Am Thorac Soc. 2023;20(12):1718–25.
56. Renzi M, Badaloni C, Trentalange A, Porta D, Davoli M, Michelozzi P. Association between air pollution, socioeconomic inequalities and cause specific mortality in a large administrative cohort in a contaminated site of central Italy. Atmos Environ. 2025;347:121082.
57. Khraishah H, Rajagopalan S. Inhaling Poor Health: The Impact of Air Pollution on Cardiovascular Kidney Metabolic Syndrome. Methodist DeBakey Cardiovasc J. 2024;20(5):47.
58. Zhang Z, Zhang G, Su B. The spatial impacts of air pollution and socio-economic status on public health: empirical evidence from China. Socio Econ Plan Sci. 2022;83:101167.
59. Wiemken TL, Kelley RR. Machine learning in epidemiology and health outcomes research. Annu Rev Public Health. 2020;41(1):21–36.

Publisher's Note