BMC Cardiovascular Disorders

RESEARCH

Open Access

Environmental chemical exposures and a machine learning-based model for predicting hypertension in NHANES 2003–2016

Kun Guo[1], Weicheng Ni[1], Leilei Du[1], Yimin Zhou[1], Ling Cheng[1] and Hao Zhou[1*]

Abstract

Background Hypertension is a common disease, often overlooked in its early stages due to mild symptoms. And persistent elevated blood pressure can lead to adverse outcomes such as coronary heart disease, stroke, and kidney disease. There are many risk factors that lead to hypertension, including various environmental chemicals that humans are exposed to, which are believed to be modifiable risk factors for hypertension.

Objective To investigate the role of environmental chemical exposures in predicting hypertension.

Methods A total of 11,039 eligible participants were obtained from NHANES 2003–2016, and multiple imputation was used to process the missing data, resulting in 5 imputed datasets. 8 Machine learning algorithms were applied to the 5 imputed datasets to establish hypertension prediction models, and the average accuracy score, precision score, recall score, and F1 score were calculated. A generalized linear model was also built to predict the systolic and diastolic blood pressure levels.

Results All 8 algorithms had good predictions for hypertension, with Support Vector Machine (SVM) being the best, with accuracy, precision, recall, F1 scores and area under the curve (AUC) of 0.751, 0.699, 0.717, 0.708 and 0.822, respectively. The R^2 of the linear model on the training and test sets was 0.28, 0.25 for systolic and 0.06, 0.05 for diastolic blood pressure.

Conclusions In this study, relatively accurate prediction of hypertension was achieved using environmental chemicals with machine learning algorithms, demonstrating the predictive value of environmental chemicals for hypertension.

Keywords Environmental chemicals, Hypertension, Machine learning, Multiple imputation

*Correspondence:
Hao Zhou
wyzh66@126.com
[1]Department of Cardiology, The First Affiliated Hospital of Wenzhou Medical University, Nanbaixiang Hospital District, Ouhai District, Wenzhou City 325000, Zhejiang Province, China

Background

Hypertension, a common cardiovascular disease [1, 2] with a prevalence of approximately 30% in adults, affects more than one billion adults worldwide [3, 4]. Persistently elevated blood pressure can lead to multiple systemic diseases such as stroke, cognitive impairment, coronary artery disease, nephropathy and retinopathy [5–8]. However, it is estimated that more than half of people with hypertension are unaware of their condition, and of those who are aware, many have poor compliance leading to inadequate treatment [9, 10]. Identifying the underlying causes can help prevent and control hypertension [11]. The etiology of hypertension involves a complex interplay of environmental and genetic susceptibilities. Among these, uncontrollable factors include age, gender, race, etc. Controllable factors include diet, physical activity level, obesity, smoking, etc.

In addition, environmental chemicals, as a class of controllable risk factors, have received increasing attention in recent years because of the inevitable exposures to them in daily life. Much evidence suggested that environmental chemicals were highly associated with hypertension. For example, Xueling Lu's review indicated that exposure to phthalates (PAEs) was a risk factor for hypertension [12]. Another paper, based on NHANES database, found that high exposure to mercury levels was associated with elevated blood pressure and prevalence of hypertension [13]. Similarly, Yao Xu found significantly higher Systolic (SBP) and diastolic blood pressure (DBP) levels in people exposed to high mixtures (Pb, Cd, Hg, As) [14]. Moreover, perfluoroalkyl substances (PFAS) was thought to be associated with elevated blood pressure, in both pregnant women and the general population [15, 16]. Polycyclic aromatic hydrocarbons (PAHs) and phenols have been associated with an increased risk of hypertension [17, 18]. Overall, a large number of studies have analyzed the association between chemical exposures and the risk of hypertension, but few have elucidated the predictive role of chemical exposures on hypertension, especially when combined with multiple categories of chemical exposures.

Machine learning, as a branch of artificial intelligence, is becoming increasingly important in the medical field due to its ability to handle large, complex and diverse data [19]. Compared to traditional predictive models, machine learning can not only handle complex data types, but also has higher accuracy, greater adaptability and intelligence [20].

The importance of early identification of hypertension cannot be overlooked, but the predictive role of environmental chemicals on hypertension has not been clarified. To explore the association between chemical exposures and hypertension, we performed an analysis by using NHANES data. In addition, machine learning-based predictive models were performed to assess the value of multiple categories of environmental chemicals in predicting hypertension.

Methods

Study population

National Health and Nutrition Examination Survey (NHANES) is a nationally representative survey conducted by the National Center for Health Statistics (NCHS) aimed at evaluating the health and nutrition status of residents in the United States. The survey is publicly available and has obtained ethical approval (https://www.cdc.gov/nchs/nhanes/irba98.htm). The survey includes 5 main categories of data: demographic, dietary, examination, laboratory, and questionnaire. In this study, we included participants from 2003 to 2016 who underwent at least one complete blood pressure measurement. Participants under 20 years of age, pregnant women, and those without urine creatinine were excluded. The environmental chemicals of interest were tested in each cycle, but not all participants were measured for all 30 environmental chemicals. Therefore, we excluded participants who were missing more than 15 environmental chemicals to ensure the accuracy and reliability of the study. In the end, a total of 11,039 participants were included in the study (Figure S1).

Data collection

Our study included participants over 7 NHANES cycles from 2003 to 2016. The public data files of the NHANES database are freely available for researchers through the NCHS website. Demographic data included age, gender, race/ethnicity, family poverty-income ratio (PIR), and education level. Dietary data included total energy intake, protein, carbohydrates, fat intake, alcohol intake, as well as potassium and sodium intake. Dietary data were obtained using 24-hour dietary recalls. Most participants had both day 1 and day 2 recall data and we calculated their average; otherwise we used data from one day. Questionnaire data included physical activity, smoking status, pack-years, and sleep disorders. Examination data included blood pressure, waist circumference, and body mass index (BMI). A total of 30 environmental chemicals were included in our study as potential predictors of hypertension. These included 4 metallic and non-metallic elements, 7 PAHs, 6 PFAS, 10 PAEs, and 3 phenols (Table S1). Detailed information on the collection of blood and urine samples and measurement methods for chemical substances could be found on the NCHS website. Values below the limits of detection (LOD), released by the national report on human exposure to environmental chemicals from the U.S. Centers for Disease Control and Prevision (CDC) were imputed as the lowest LOD divided by the square root of two (NHANES 2009–2010:

Phthalates - Urine Data Documentation, Codebook, and Frequencies (cdc.gov)). As recommended, urinary creatinine was used as a reference to account for urinary dilution [21].

Data pre-processing

First of all, environmental exposures with missing values greater than 50% were excluded from the study, including 6 PFAS and As, which could have significant implication on data analysis. Subsequently, a natural logarithm transformation was applied to other 23 environmental chemicals to improve their normal distribution. Missing values were then imputed with multiple imputation, which is an effective and widely accepted method for handling missing values [22, 23]. We performed the 'pmm' methods in the 'mice' R package to impute missing values and obtained 5 slightly different imputed datasets, meaning that their values were not exactly the same, but not statistically different. The 'Predictive Mean Matching' (PMM) algorithm matches the missing value with several non-missing observed values, and then randomly selects one value from these most matched non-missing values to impute, which was used to handle most missing values. To reduce selection bias, participants were assigned to the same training/test fold (70:30) in the 5 imputed datasets.

Diagnostic criteria for hypertension and definition of some covariates

Participants were diagnosed with hypertension if any of the following conditions were met. Mean systolic blood pressure ≥ 140 mmHg, mean diastolic blood pressure ≥ 90 mmHg, self-reported hypertension, and participants taking antihypertensive medication. The mean blood pressure was the average of at least 1 and at most 3 blood pressure values.

Some covariates were defined as described below. Smoking status was categorized as never smokers versus ever smokers. Pack-years, calculated for ever smokers only, were defined as: (packs per day) × (years smoked). Physical activity was calculated on the basis of minutes of exercise per week [24] and was classified as ideal(≥ 75 min of vigorous activity or ≥ 150 min of moderate activity per week), intermediate(1–74 min of vigorous activity or 1–149 min of moderate activity per week) and poor(0 min of moderate or vigorous activity per week) [25]. Sleep disorder was defined as "Ever been told by a doctor or other health professional that you have a sleep disorder".

Regularized partial correlation network

Regularized partial correlation network was used to measure the degree of association between two environmental chemicals while controlling for the influence of other chemicals [26]. The nodes of this network represented different environmental chemicals, and the edge weights represented the partial correlation coefficients. Blue edges indicated positive correlations, while red edges indicated negative correlations.

Machine learning algorithms

8 machine learning algorithms were applied in this study, including Generalized Linear Model with Elastic Net Regularization (GLMNET), Support Vector Machine (SVM), Random Forest (RF), Kernel K-Nearest Neighbor (KKNN), Random Tree (RT), Neural Network (NN), Naive Bayes (NB) and Tree bag (TB). All of the 41 characteristics were included in the prediction model, including 23 environmental chemicals and 18 other covariates. In the training set, each algorithm performed a grid search and 5-fold cross-validation to select the best independent variables and hyperparameters. Then, the test set was used to evaluate the generalization ability and performance of the models. According to the proposal, we calculated the average accuracy, precision, recall, F1 score and AUC for the 5 imputed datasets to measure the performance of the classifiers.

In addition, we used the 'caret' package in R to evaluate the contribution of each variable to the predictive model. Specifically, we applied the 'varImp' function, which automatically calculates importance scores based on the impact of each variable on the model's predictions. Variables with higher scores contribute more to the model. These scores were then used to create bar charts that visually display the relative importance of each variable.

Identify characteristics that affect blood pressure

When analyzing factors that influence blood pressure levels, all participants taking antihypertensive medication were excluded due to the significant effect of these drugs on blood pressure. Systolic and diastolic blood pressure values were separated by percentiles to investigate whether the intensity of environmental exposure changed with different blood pressure levels. Principal component analysis (PCA) was used to investigate the association between environmental chemicals and blood pressure by converting the raw high-dimensional data into two dimensions, with the first two principal components (PC1 and PC2) visualising the relative positions and distributions between the different blood pressure groups. Permutational multivariate analysis of variance (PERMANOVA) was used to test the statistical significance of the association between chemical distribution and levels of SBP and DBP.

In addition, linear regression was used to identify environmental chemicals that changed significantly with changes in blood pressure (ln-transformed intensity). All other factors included in the analysis were adjusted for as

covariates, including gender, age, ethnicity, family PIR, education level, BMI, diet, smoking, pack-years, alcohol intake, sleep disturbance and physical activity level.

Statistic analysis

All statistics were performed with R version 4.2.3 and a two-tailed $p<0.05$ was considered statistically different.

Result

Baseline characteristics of participants

Our study ultimately included 11,039 participants, of whom 4739 (43%) were diagnosed with hypertension. 2978 (63%) of the hypertensive patients were taking antihypertensive drugs. The mean SBP and DBP were 121.3 (111.3, 134.0) mmHg and 70.7 (62.7, 77.3) mmHg, respectively. The general characteristics and the distribution of the environmental chemicals can be viewed in Table 1. The relative stability of chemical concentrations over the years is shown in Figure S2. Moreover, multiple imputation performed well (Table S2) and there was no statistical difference between the characteristics of the original data and the imputed datasets ($p>0.05$).

Correlation network among environmental chemicals

We performed a regularized biased correlation network with the 23 environmental chemicals mentioned above. As we can see (Fig. 1), most of the environmental chemicals in the same category clustered together and were positively correlated, suggesting that the chemicals may act as a group rather than individually. Most of the chemicals showed low correlation, for example, cadmium and lead ($r=0.24$, $P<0.05$), while a few exhibited high correlation, such as P03 and P04 ($r=0.75$, $P<0.05$).

Performance of machine learning model for predicting hypertension

According to the results shown in Table 2, GLMNET and SVM were able to predict hypertension relatively accurately with AUCs of 0.821 and 0.822, respectively, while NB wasless effective with AUCs of 0.769.

Taking both AUC and characteristic importance into account, the SVM algorithm performed best. The accuracy, precision, recall and F1 scores of SVM were 0.751, 0.699, 0.717 and 0.708, respectively. Figure 2A showed the top 12 contributors to the model. Among them, the top 3 contributors were age, waist circumference and BMI, while the chemicals that contributed most to the prediction of hypertension were Pb, P10, and MHP. We also found that the total contribution of each chemical category played a relatively important role in the model. PAEs, PAHs, elements and phenols accounted for 5.8%, 6.1%, 6.9%, 2.6% of the total contribution, respectively (Fig. 2C). In addition, the AUCs of the training and test sets were 0.819 and 0.822, respectively (Fig. 2B), indicating that the SVM model achieved relatively good prediction of hypertension without overfitting

Associations between environmental chemicals and blood pressure

PCA showed association between chemicals and SBP, DBP. And PERMANOVA revealed a statistically significant difference in the change in intensity of environmental chemicals as SBP and DBP levels changed ($p<0.01$ for both SBP and DBP) (Fig. 3A-B). Linear regression was used to further investigate the environmental chemicals that changed most significantly with blood pressure levels. Figure 3 C-D showed that MHH, P25 and MNP were positively associated with changes in SBP; MHH, P06, Cd and Pb were positively associated with changes in DBP.

Linear regression models were developed to predict SBP and DBP levels, but the model performed poorly. The R^2 for predicting SBP was 0.28 and 0.25 in the training and test sets, respectively, and the R^2 for predicting DBP was 0.06, 0.05, respectively (Fig. 4).

Discussion

Hypertension is the most important modifiable risk factor for all-cause mortality worldwide [9]. Investigation of the underlying risk factors for hypertension can be of great help in the prevention and control of the disease. In this study, we investigated the association between environmental chemicals and hypertension, identifying several significantly relevant environmental chemicals (e.g., Pb). Furthermore, with a machine learning-based prediction model, we not only achieved relatively accurate predictions of hypertension (binary outcome), but also identified significant contributions of environmental chemicals to the prediction of hypertension. However, the linear model was not a satisfactory predictor for continuous outcomes (systolic and diastolic blood pressure).

Exposure to environmental chemicals is unavoidable as they are commonly found in air, water and daily supplies. Many studies have demonstrated the relevance of environmental chemicals and hypertension, in both adults and children [27, 28]. Nevertheless, only a limited number of studies have examined the predictive value of these factors for hypertension [29, 30]. The majority of these studies have focused on the correlation between a single chemical element or a class of chemical elements and hypertension, which can increase the risk of hypertension [31]. Nevertheless, it remains unclear whether the combined effect of multiple type of environmental chemicals in the human body will be additive, antagonistic, or whether they will exert their influence on blood pressure individually.

Our study focused on the prediction of hypertension using multiple environmental chemicals. Therefore, we built 8 machine learning-based models, most

Table 1 Characteristics of the total participants

Characteristics	Total	Training set (*n* = 7728)	Test set (*n* = 3311)	*P* value
Hypertension (%)				0.26
no	6300 (57)	4383 (57)	1917 (58)	
yes	4739 (43)	3345 (43)	1394 (42)	
Current use of antihypertensive drugs (%)				
no	8061 (73)			
yes	2978 (27)			
SBP (mmHg)	121.3 (111.3, 134.0)	121.3 (111.33, 134.0)	122.0 (111.3, 134.0)	0.67
DBP (mmHg)	70.7 (62.7, 77.3)	70.7 (62.7, 77.3)	70.0 (62.7, 77.3)	0.34
All covariates				
Gender (%)				0.35
male	5515 (50)	3838 (50)	1677 (51)	
female	5524 (50)	3890 (50)	1634 (49)	
Race or ethnicity (%)				0.42
Hispanic	2741 (25)	1894 (25)667 (9)	847 (26)	
Non-Hispanic white	4894 (44)	3434 (44)	1460 (44)	
Non-Hispanic Black	2373 (21)	1687 (22)	686 (21)	
Other race	1031 (9)	713 (9)	318 (10)	
Education level (%)				0.25
Less than 9th grade	1234 (11)	843 (11)	391 (12)	
9–11th grade	1635 (15)	1150 (15)	485 (15)	
High school graduate	2578 (23)	1787 (23)	791 (24)	
Some college or AA degree	3168 (29)	2212 (29)	956 (29)	
College graduate or above	2424 (22)	1736 (22)	688 (21)	
Physical activity (%)				0.72
poor	6240 (57)	4381 (57)	1859 (56)	
intermediate	1391 (13)	979 (13)	412 (12)	
ideal	3408 (31)	2368 (31)	1040 (31)	
Smoking (%)				0.31
never	5957 (54)	4195 (54)	1762 (53)	
ever	5082 (46)	3533 (46)	1549 (47)	
Smoking pack-years	0 (0, 7.5)	0 (0, 7)	0 (0, 8.65)	0.06
Sleep disorders, n (%)				0.32
yes	2735 (25)	1936 (25)	799 (24)	
no	8304 (75)	5792 (75)	2512 (76)	
Age(years)	49 (34, 64)	49 (34, 64)	49 (34, 64)	0.79
Family PIR	2.1 (1.1, 4.1)	2.1 (1.1, 4.1)	2.1 (1.1, 4.1)	0.33
BMI (kg/m^2)	28.0 (24.3, 32.4)	28.0 (24.3, 32.4)	28.1 (24.3, 32.6)	0.97
Waist(cm)	97.8 (87.7, 108.9)	97.9 (87.6, 108.9)	97.5 (87.9, 108.8)	0.86
Total energy intake (kcal)	1908.5 (1470.75, 2500.5)	1915 (1479, 2504.12)	1891.5 (1452.75, 2480.5)	0.20
potassium intake(mg)	2458 (1855, 3164)	2457.5 (1856.25, 3171.5)	2458.5 (1853.75, 3147.25)	0.81
sodium intake(mg)	3123 (2321.5, 4097)	3129.5 (2329, 4089.62)	3103.5 (2289.5, 4102.75)	0.28
protein intake(g)	74.61 (55.71, 98.14)	74.49 (56.1, 98.53)	74.81 (54.84, 97.16)	0.16
Carbohydrate intake(g)	232.57 (174.39, 305.19)	232.49 (175.46, 304.11)	232.61 (172.4, 307.63)	0.64
Fat intake (g)	70.6 (50.23, 97.24)	70.84 (50.74, 97.39)	69.79 (48.76, 96.51)	0.09
Alcohol intake (g)	0 (0, 7)	0 (0, 6.48)	0 (0, 7)	0.24
Environmental chemicals				
Hg_log (ug/L)	-0.11 (-0.73, 0.58)	-0.11 (-0.73, 0.58)	-0.12 (-0.73, 0.58)	0.87
Pb_log (ug/dL)	0.27 (-0.17, 0.74)	0.27 (-0.16, 0.74)	0.28 (-0.17, 0.73)	0.78
Cd_log (ug/L)	-1.05 (-1.51, -0.49)	-1.08 (-1.56, -0.49)	-1.05 (-1.51, -0.46)	0.34
P01_log (ng/L)	7.57 (6.67, 8.83)	7.58 (6.68, 8.84)	7.55 (6.65, 8.79)	0.6
P02_log (ng/L)	8.26 (7.43, 9.14)	8.26 (7.41, 9.15)	8.26 (7.46, 9.12)	0.69
P03_log (ng/L)	4.4 (3.69, 5.53)	4.41 (3.69, 5.54)	4.39 (3.69, 5.52)	0.26

Table 1 (continued)

Characteristics	Total	Training set (*n* = 7728)	Test set (*n* = 3311)	*P* value
P04_log (ng/L)	5.43 (4.74, 6.39)	5.44 (4.74, 6.39)	5.4 (4.74, 6.38)	0.34
P06_log (ng/L)	4.86 (4.25, 5.5)	4.87 (4.25, 5.5)	4.85 (4.24, 5.49)	0.44
P10_log (ng/L)	4.57 (3.87, 5.33)	4.57 (3.87, 5.34)	4.55 (3.83, 5.31)	0.49
P25_log (ng/L)	4.94 (4.3, 5.63)	4.95 (4.3, 5.64)	4.93 (4.32, 5.62)	0.7
BPA_log (ng/mL)	0.47 (-0.22, 1.16)	0.47 (-0.22, 1.16)	0.47 (-0.36, 1.16)	0.42
BP3_log (ng/mL)	2.52 (1.34, 4.06)	2.51 (1.31, 4.09)	2.53 (1.36, 4.01)	0.96
TCS_log (ng/mL)	2.07 (0.53, 3.84)	2.1 (0.59, 3.9)	2 (0.49, 3.74)	0.08
MBP_log (ng/mL)	2.68 (1.89, 3.38)	2.68 (1.87, 3.36)	2.68 (1.92, 3.42)	0.11
MEP_log (ng/mL)	4.23 (3.18, 5.38)	4.25 (3.17, 5.37)	4.2 (3.18, 5.42)	0.69
MHP_log (ng/mL)	0.41 (-0.42, 1.28)	0.41 (-0.37, 1.25)	0.41 (-0.49, 1.31)	0.51
MNP_log (ng/mL)	-0.14 (-0.45, 0.09)	-0.14 (-0.45, 0.09)	-0.14 (-0.45, 0.09)	0.69
MZP_log (ng/mL)	1.74 (0.83, 2.58)	1.74 (0.83, 2.59)	1.74 (0.86, 2.57)	0.76
MC1_log (ng/mL)	0.74 (-0.11, 1.52)	0.74 (-0.11, 1.53)	0.74 (-0.11, 1.5)	0.66
MHH_log (ng/mL)	2.43 (1.63, 3.22)	2.42 (1.61, 3.2)	2.46 (1.65, 3.27)	0.13
MOH_log (ng/mL)	1.96 (1.16, 2.73)	1.95 (1.16, 2.72)	1.99 (1.19, 2.77)	0.13
MIB_log (ng/mL)	1.96 (1.19, 2.64)	1.95 (1.16, 2.66)	2 (1.25, 2.62)	0.2
ECP_log (ng/mL)	2.86 (2.1, 3.63)	2.85 (2.1, 3.61)	2.88 (2.12, 3.68)	0.18

Notes: normally distributed continuous variables: mean±standard deviation; non-normally distributed continuous variables: median (interquartile range); categorical variables: percentages. And the numerical values of all chemical elements were represented using logarithmic values with base e

Abbreviations: SBP: systolic blood pressure; DBP: diastolic blood pressure; PIR: poverty-income ratio; BMI: body max index

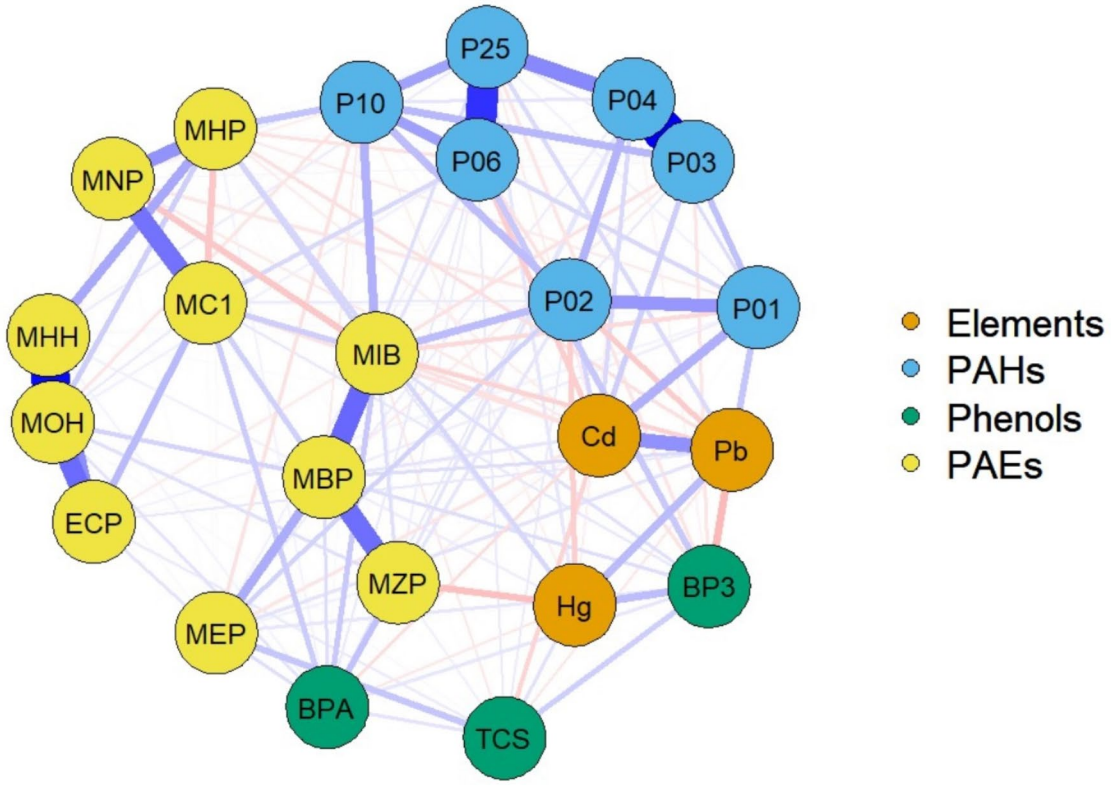

Fig. 1 Regularized partial correlation network. Edge weights represented partial correlation coefficients. Blue edges indicated positive correlations, while red edges indicated negative correlations

Table 2 Performance of 8 algorithms to predict hypertension

Algorithms	Accuracy	Precision	Recall	F1 score	AUC
SVM	0.751	0.699	0.717	0.708	0.822
GLMNET	0.754	0.711	0.702	0.706	0.821
NN	0.751	0.695	0.727	0.711	0.819
RF	0.753	0.703	0.715	0.709	0.810
TB	0.742	0.693	0.695	0.694	0.799
KKNN	0.725	0.723	0.561	0.632	0.792
NB	0.716	0.646	0.723	0.682	0.769
RT	0.743	0.697	0.689	0.693	0.793

All values above were the average of each algorithm over 5 imputed datasets

Abbreviations: AUC: area under curve; SVM: support vector machine; GLMNET: generalized linear model with elastic-net regularization; NN: neural network; RF: random forest; TB: tree bag; KKNN: kernel K-nearest neighbor algorithm; NB: naive bayes; RT: random tree

of which achieved good predictions, demonstrating the high potential and feasibility of machine learning for hypertension prediction. Then, taking into account the performance and interpretability of the models, the SVM was selected for further analysis.

On the basis of model contribution, the top 3 environmental chemicals for predicting hypertension were Pb, P10, and MHP. The possible mechanisms by which they cause hypertension are discussed below. Previous studies have shown that Pb causes hypertension by inducing vasoconstriction (alpha adrenergic receptors) and regulating the production of renin and angiotensin [32]. Besides, prolonged exposure to PAH is associated with oxidative stress. The resulting vasoconstriction and endothelial cell dysfunction are thought to be the mechanism responsible for the increase in blood pressure [33]. In the case of phthalates, studies have found that they may affect blood pressure through a variety of mechanisms. These include impairment of endothelial function and interference with the renin-angiotensin system [34, 35]. The above studies suggest that environmental chemicals

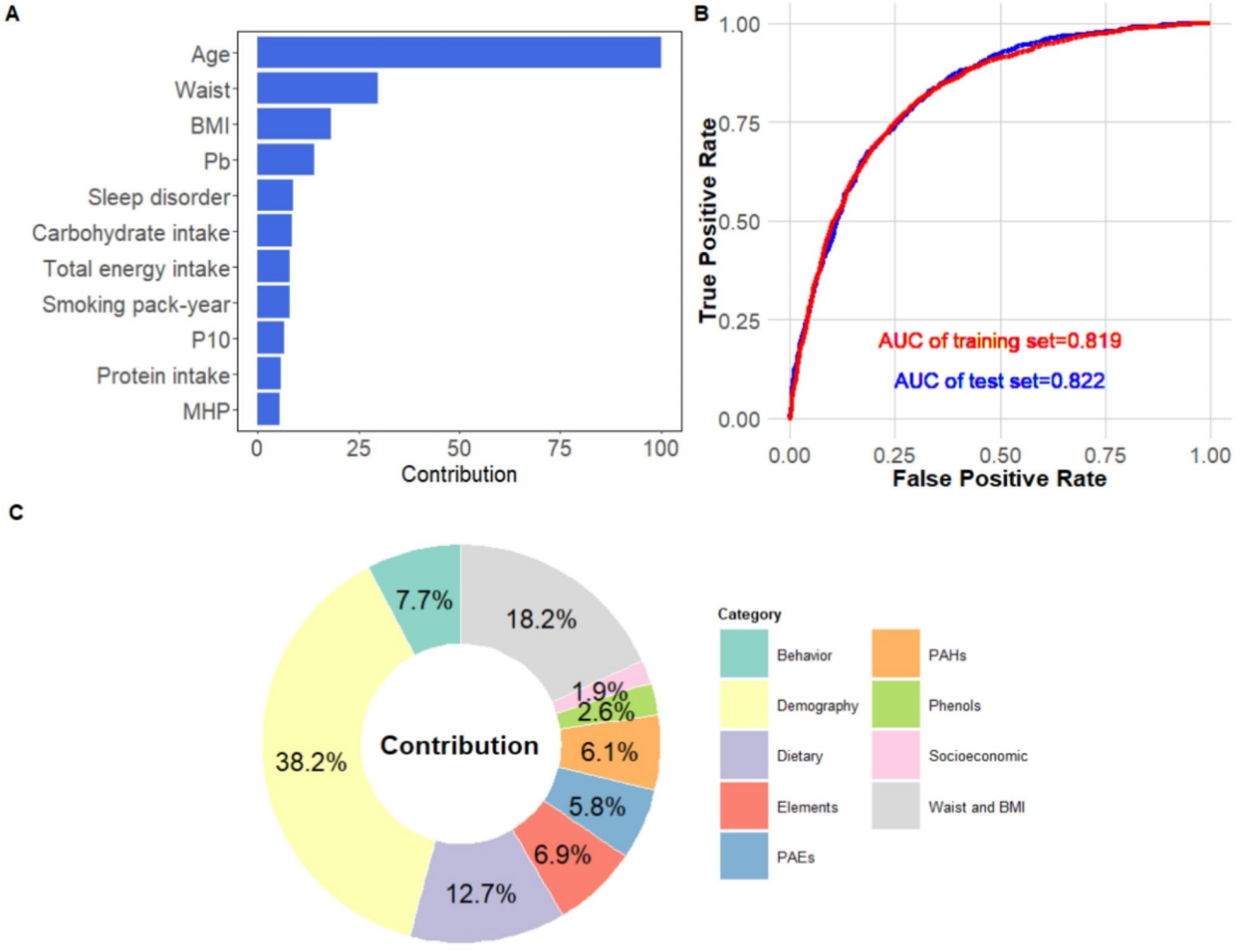

Fig. 2 Performance of the SVM model. (**A**) Contribution of characteristics to the prediction of hypertension. (**B**) AUC of training and test sets for hypertension prediction in SVM model. (**C**) Contribution of each category of characteristics to the prediction of hypertension

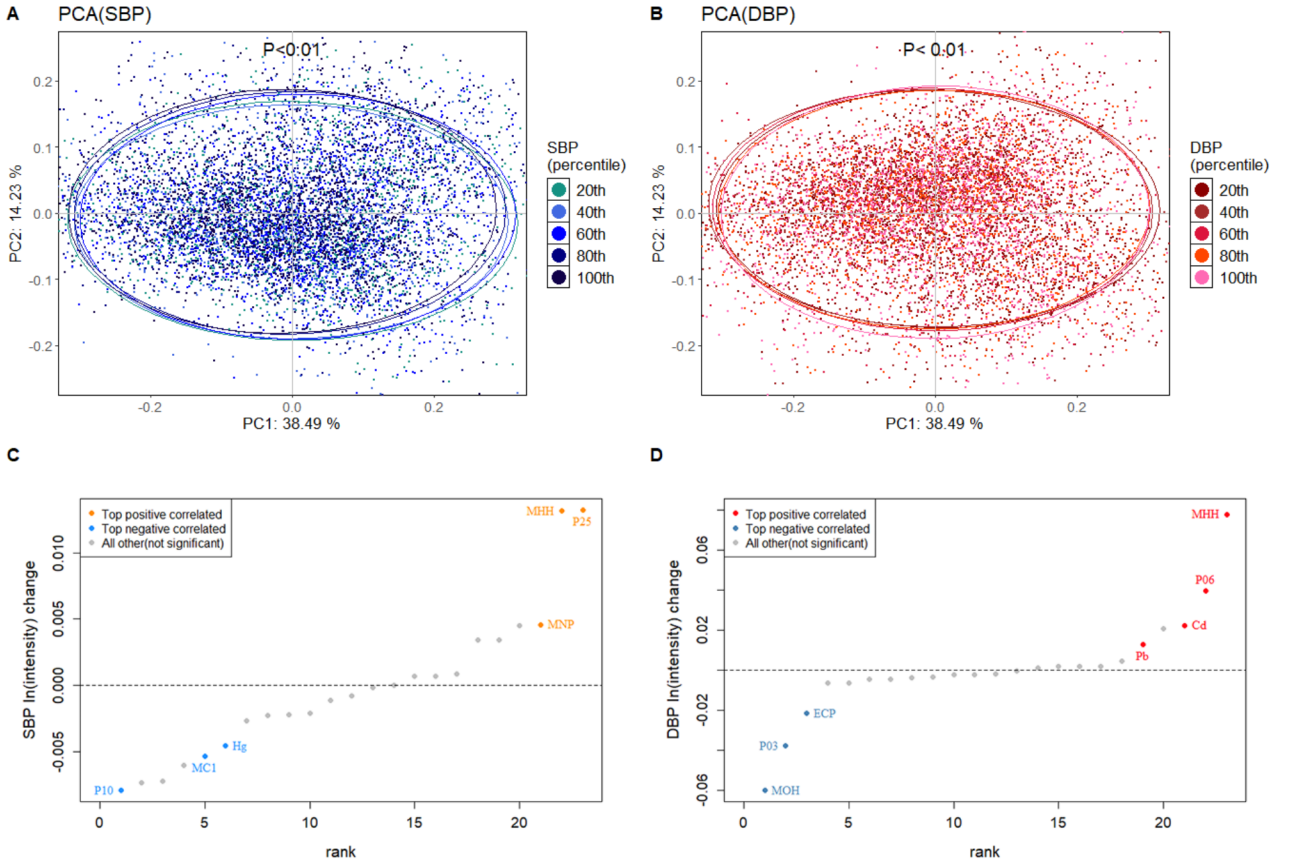

Fig. 3 Association of systolic and diastolic blood pressure with environmental chemicals. (**A**-**B**) PCA was used to assess the association between elevated blood pressure and environmental chemicals. (**C**-**D**) Linear regression to identify chemicals significantly associated with elevated blood pressure ($P < 0.05$)

are involved in the progression of hypertension and partially illustrate the predictive value of environmental chemicals for hypertension.

In addition, our study found that there was often a positive correlation between chemicals in the same category, such as Pb and Cd. It is possible that they have a synergistic effect on blood pressure, but this needs to be investigated further. Similarly, Juhua Luo's previous study showed that metal mixtures had a greater effect on kidney function than single metal [36].

In line with previous studies, traditional risk factors also made a significant contribution to the prediction of hypertension [37]. These included the non-modifiable risk factors: sex, age and race, and the modifiable risk factors: dietary, behavior, waist circumference and BMI. In addition, we also found that the combination of BMI and waist circumference was a better predictor of obesity-related hypertension [38].

In another place, when the study outcome was changed from hypertension or not to systolic and diastolic blood pressure, we also excluded patients taking antihypertensive medication. This may have led to a change in the environmental chemicals affecting the two outcomes, but we felt it was necessary because of the dramatic effect of antihypertensive medication on blood pressure. This may partly explain why the three most contributing environmental chemicals in the SVM model, Pb, and P10, also had an effect on SBP or DBP in the linear regression, whereas MHP had no effect on them.

Our study had several strengths. First, we applied multiple imputation to deal with missing values, which was a widely accepted and effective approach. Multiple imputation allows for the full use of available data, reducing bias caused by missing values and improving the reliability of the results. Second, we used several categories of environmental chemicals to predict hypertensive individuals, which provided a new perspective on hypertension prediction. Finally, we performed several machine learning-based prediction models, and most of them achieved good hypertension prediction.

However, there were some limitations of this study. First, our sample coverage may have been limited, which may have prevented us from capturing the effects of environmental pollution on hypertension in certain areas or in certain populations. Second, this

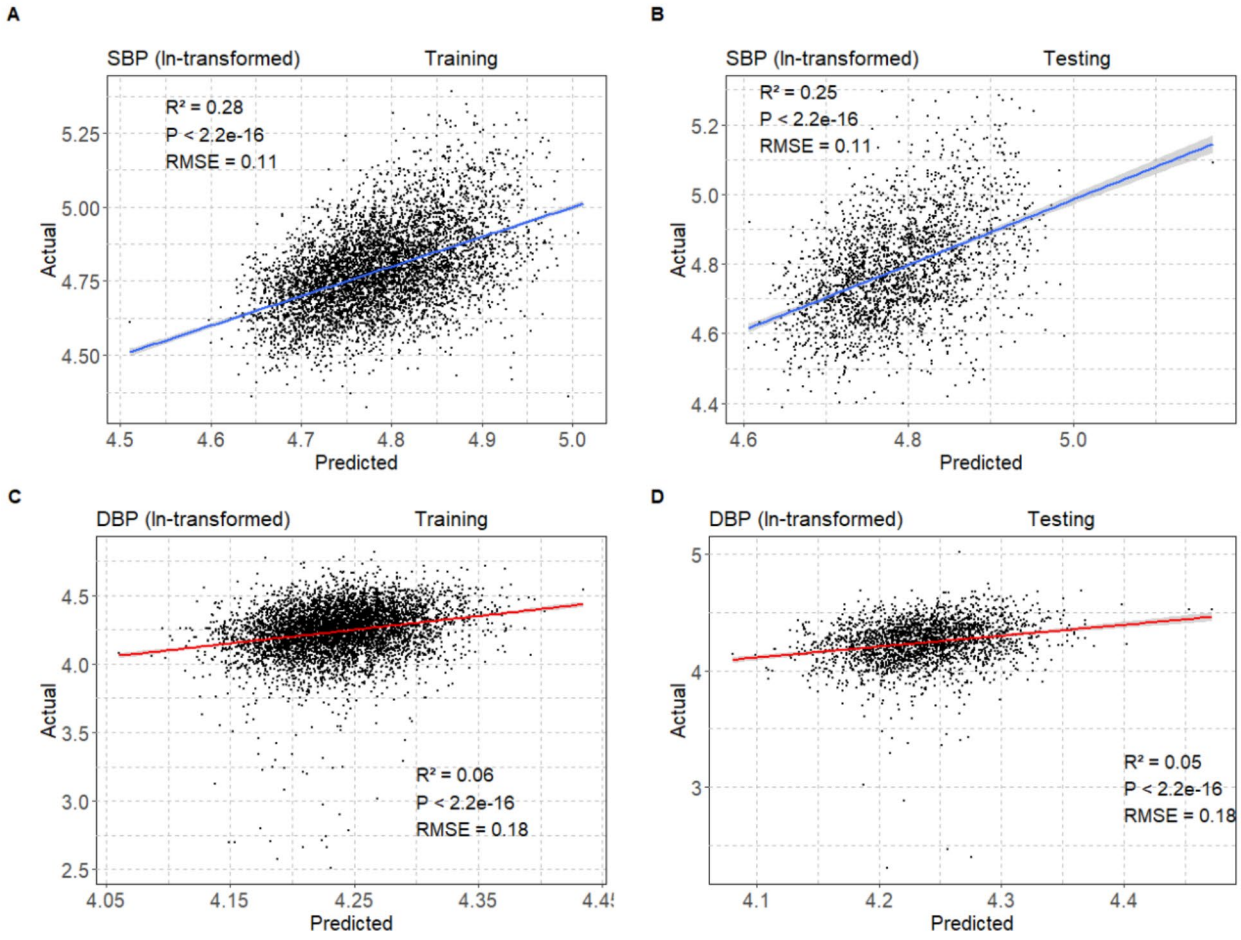

Fig. 4 Linear model predicted systolic and diastolic blood pressure. (**A**-**B**) Performance of a linear model for predicting SBP on the training and test sets. (**C**-**D**) Performance of a linear model for predicting DBP on the training and test sets
Abbreviations: SBP: systolic blood pressure; DBP: diastolic blood pressure; RMSE: root mean square error

was a cross-sectional study and could not really explain the long-term effects of environmental chemical exposure on individual blood pressure, which needs to be confirmed by further cohort studies. Third, due to database limitations, the effect of genetic factors (e.g. family history of hypertension) on blood pressure was not taken into account. Finally, the multiple imputation model was applied to the entire dataset to handle missing values, including both the training and testing sets. While this approach enhances data completeness, it may also introduce the potential risk of information leakage, potentially leading to an overestimation of the model's performance.

Conclusion

By analysing NHANES data from 2003 to 2016, the machine learning-based prediction model showed that environmental chemicals could predict hypertension with relative accuracy. The SVM algorithm showed that Pb, P10, and MHP were the main environmental factors predicting hypertension. Our study suggests that the role of environmental chemical exposures in hypertension cannot be ignored.

Abbreviations

BMI	Body mass index
CDC	Centers for Disease Control and Prevision
DBP	Diastolic blood pressure
GLMNET	Generalized Linear Model with Elastic Net Regularization
KKNN	Kernel K-Nearest Neighbor
LOD	Limits of detection
NB	Naive Bayes
NCHS	National Center for Health Statistics
NHANES	National Health and Nutrition Examination Survey
NN	Neural Network
PAEs	Phthalates
PAHs	Polycyclic aromatic hydrocarbons
PCA	Principal component analysis
PERMANOVA	Permutational multivariate analysis of variance
PFAS	Perfluoroalkyl substances
PIR	Poverty-income ratio
PMM	Predictive Mean Matching
RF	Random Forest
RMSE	Root mean square error
RT	Random Tree

SBP	Systolic blood pressure
SVM	Support Vector Machine
TB	Tree bag

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12872-024-04216-z.

Supplementary Material 1

Supplementary Material 2

Supplementary Material 3

Supplementary Material 4

Acknowledgements
Not applicable.

Author contributions
K.G. drafted the work. K.G., WC.N., and H.Z. were responsible for the conception and design of the work and contributed to formal analysis. K.G. and LL.D. handled visualization, while H.Z. supervised the project. YM.Z. and L.C. Involved in data acquisition and curation. K.G., LL.D. and YM.Z. reviewed and edited the manuscript. All authors reviewed the manuscript.

Funding
This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability
The datasets generated and/or analyzed during the current study are available in the website, [https://wwwn.cdc.gov/nchs/nhanes/Default.aspx].

Declarations

Ethics approval and consent to participate
The survey data used in this study are publicly available and have received ethical approval. Detailed information regarding the ethical review and approval can be found at " https://www.cdc.gov/nchs/nhanes/irba98.htm ".

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 14 June 2024 / Accepted: 20 September 2024
Published online: 09 October 2024

References

1. Fernández-Ruiz I. Systolic and diastolic hypertension independently predict CVD risk. Nat Rev Cardiol. 2019;16:578–9.
2. Ansah JP, et al. An evaluation of the impact of aggressive hypertension, diabetes and smoking cessation management on CVD outcomes at the population level: a dynamic simulation analysis. BMC Public Health. 2019;19:1105.
3. Carey RM, et al. Treatment of hypertension: a review. JAMA. 2022;328:1849–61.
4. Mills KT, et al. The global epidemiology of hypertension. Nat Rev Nephrol. 2020;16:223–37.
5. Cheung CY, et al. Hypertensive eye disease. Nat Rev Dis Primers. 2022;8:14.
6. Anderson AH, et al. Time-updated systolic blood pressure and the progression of chronic kidney disease: a cohort study. Ann Intern Med. 2015;162:258–65.
7. Ungvari Z, et al. Hypertension-induced cognitive impairment: from pathophysiology to public health. Nat Rev Nephrol. 2021;17:639–54.
8. Cressman MD, Gifford RW. Jr. Hypertension and stroke. J Am Coll Cardiol. 1983;1:521–7.
9. Oparil S, et al. Hypertension. Nat Rev Dis Primers. 2018;4:18014.
10. Burnier M, Egan BM. Adherence in hypertension. Circ Res. 2019;124:1124–40.
11. Cooper R. Hypertension, genes, and environment: challenges for prevention and risk prediction. Circulation. 2018;137:662–4.
12. Lu X, et al. Phthalate exposure as a risk factor for hypertension. Environ Sci Pollut Res Int. 2018;25:20550–61.
13. Tang J, et al. Total arsenic, dimethylarsinic acid, lead, cadmium, total mercury, methylmercury and hypertension among Asian populations in the United States: NHANES 2011–2018. Ecotoxicol Environ Saf. 2022;241:113776.
14. Yao X, et al. Stratification of population in NHANES 2009–2014 based on exposure pattern of lead, cadmium, mercury, and arsenic and their association with cardiovascular, renal and respiratory outcomes. Environ Int. 2021;149:106410.
15. Bao WW, et al. Gender-specific associations between serum isomers of perfluoroalkyl substances and blood pressure among Chinese: isomers of C8 health project in China. Sci Total Environ. 2017;607–608:1304–12.
16. Preston EV, et al. Early-pregnancy plasma per- and polyfluoroalkyl substance (PFAS) concentrations and hypertensive disorders of pregnancy in the project viva cohort. Environ Int. 2022;165:107335.
17. Lu L, Ni R. Association between polycyclic aromatic hydrocarbon exposure and hypertension among the U.S. adults in the NHANES 2003–2016: a cross-sectional study. Environ Res. 2023;217:114907.
18. Jiang S, et al. Association of bisphenol A and its alternatives bisphenol S and F exposure with hypertension and blood pressure: a cross-sectional study in China. Environ Pollut. 2020;257:113639.
19. Handelman GS, et al. eDoctor: machine learning and the future of medicine. J Intern Med. 2018;284:603–19.
20. Ngiam KY, Khor IW. Big data and machine learning algorithms for health-care delivery. Lancet Oncol. 2019;20:e262–73.
21. Barr DB, et al. Urinary creatinine concentrations in the U.S. population: implications for urinary biologic monitoring measurements. Environ Health Perspect. 2005;113:192–200.
22. Allotey PA, Harel O. Multiple imputation for incomplete data in environmental epidemiology research. Curr Environ Health Rep. 2019;6:62–71.
23. Chevret S, et al. Multiple imputation: a mature approach to dealing with missing data. Intensive Care Med. 2015;41:348–50.
24. Ladabaum U, et al. Obesity, abdominal obesity, physical activity, and caloric intake in US adults: 1988 to 2010. Am J Med. 2014;127:717–e727712.
25. Booth JN 3rd, et al. Trends in prehypertension and hypertension risk factors in US adults: 1999–2012. Hypertension. 2017;70:275–284.
26. Epskamp S, Fried EI. A tutorial on regularized partial correlation networks. Psychol Methods. 2018;23:617–34.
27. Qu Y, et al. Effect of exposures to mixtures of lead and various metals on hypertension, pre-hypertension, and blood pressure: a cross-sectional study from the China National Human Biomonitoring. Environ Pollut. 2022;299:118864.
28. Warembourg C, et al. Early-life environmental exposures and blood pressure in children. J Am Coll Cardiol. 2019;74:1317–28.
29. Li W, et al. Effects of heavy metal exposure on hypertension: a machine learning modeling approach. Chemosphere. 2023;337:139435.
30. Peters JL, et al. Epidemiologically-informed cumulative risk hypertension models simulating the impact of changes in metal, organochlorine, and non-chemical exposures in an environmental justice community. Environ Res. 2019;176:108544.
31. Zhou S, et al. Paraben exposures and their interactions with ESR1/2 genetic polymorphisms on hypertension. Environ Res. 2022;213:113651.
32. Mitra P, et al. Clinical and molecular aspects of lead toxicity: an update. Crit Rev Clin Lab Sci. 2017;54:506–28.
33. He J, et al. Environmental dose of 16 priority-controlled PAHs mixture induce damages of vascular endothelial cells involved in oxidative stress and inflammation. Toxicol Vitro. 2022;79:105296.
34. Rahmani A, et al. Prenatal exposure to phthalic acid induces increased blood pressure, oxidative stress, and markers of endothelial dysfunction in rat offspring. Cardiovasc Toxicol. 2016;16:307–15.
35. Jaimes R 3, et al. Plastics and cardiovascular health: phthalates may disrupt heart rate variability and cardiovascular reactivity. Am J Physiol Heart Circ Physiol. 2017;313:H1044–53.
36. Luo J, Hendryx M. Metal mixtures and kidney function: an application of machine learning to NHANES data. Environ Res. 2020;191:110126.
37. Zhao H, et al. Predicting the risk of hypertension based on several easy-to-collect risk factors: a machine learning method. Front Public Health. 2021;9:619429.

38. Zhang M, et al. Body mass index and waist circumference combined predicts obesity-related hypertension better than either alone in a rural Chinese population. Sci Rep. 2016;6:31935.

Publisher's note