Appendixes

Appendix A. Assignment of variables

Table S1 Assignment of variables.

Variables	Assigning values
Body mass index (BMI)	0, BMI < 18.5 kg/m^2
	1, 18.5 kg/m^2 ≤ BMI < 25 kg/m^2
	2, BMI ≥ 25 kg/m^2
Waist-to-hip ratio (WHR)	0, WHR < 0.9 for male, WHR < 0.8 for female
	1, WHR ≥ 0.9 for male, WHR ≥ 0.8 for female
Population density of the district*	0, Below the average of Beijing
	1, Equal to or above the average of Beijing
Per capita disposable income of the district*	0, Below the medium level of Beijing
	1, Equal to or above the medium level in Beijing

Notes: *, the average population density and the medium level of per capita disposable income of Beijing, was from the official website of the Beijing Municipal Bureau of Statistics (China).

Appendix B. Exposure assessment

In this study, generalized additive models (GAM), land use regression model (LUR) and back propagation neural network (BPNN) were used to assess the annual exposure levels of air pollution for each participant. The concentrations of 6 air pollutants collected at 35 monitoring points were used as the dependent variables, and geographical environment elements, population density and meteorological factors within the buffer zone at different distances around the stations were used as the independent variables. Among them, geographical environment elements included the length of roads and the area occupied by each land use types. The data of roads were obtained from the public vector road network map on the OpenStreetMap website. Land use types include five major categories: building land, arable land, forest land, grassland and water, with data obtained from the Resource and Environment Science Data Centre of the Chinese Academy of Sciences. The total length of roads within 100 m, 300 m and 500 m radius buffer zones at the center of each monitoring point or workplace, as well as the length of roads and the area of each land use type within 1 km, 2 km, 3 km and 4 km radius buffer zones were calculated. The data of population density were obtained from the Beijing Municipal Bureau of Statistics. Meteorological factors included temperature, relative humidity, air pressure, wind speed and wind index.

Observations of the first four meteorological factors at the 35 monitoring stations were obtained from the China Meteorological Administration. The corresponding data of each workplace were assessed through the Kriging interpolation (Lee et al., 2012). According to the wind rose map of Beijing, the wind direction of south-southwest (SSW) was selected as main direction, which had the highest frequency. And the wind index of each monitoring point and workplace was calculated according to Eq. (S1).

$$Wind\ index = \left[1 - \cos\left(\theta - SSW\right)\right]/2 \quad , \qquad \text{(S1)}$$

where, θ is the angle from the nearest air pollution industrial source to the monitoring point or workplace. The wind index takes a value of 0–1, where 0 denotes that the site is upwind of the nearest source, and 1 denotes that it is downwind. Data on industrial sources were obtained from the national list of key monitoring enterprises, issued by the Ministry of Environmental Protection.

Considering the nonlinear relationship between meteorological factors and air pollutants, GAM was used to fit the LUR, and a spline smoothing function was introduced to adjust the meteorological factors in the model. We iterated through all the variables by using the algorithm of the European Study of Cohorts to Air Pollution Effects, which eventually allowed the model to converge with a largest R^2 (Hoek et al., 2011). The variables in the final model were incorporated into the BPNN to build a model for predicting air pollutant concentrations. The accuracy of the model was tested using the 10-fold cross-validation method. The effectiveness of the model was measured by mean absolute error (MAE), mean absolute percentage error (MAPE), root mean square error (RMSE) and degree of accuracy (D).

$$MAE = \frac{1}{n}\sum_{t=1}^{n}\left(\left|Y_t - y_t\right|\right) \quad , \qquad \text{(S2)}$$

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left(\left|Y_t - y_t\right|/\left|Y_t\right|\right) \quad , \qquad \text{(S3)}$$

$$RMSE = \left[\frac{1}{n}\sum_{t=1}^{n}\left(Y_t - y_t\right)^2\right]^{1/2} \quad , \qquad \text{(S4)}$$

$$D = \left[1 - \frac{MAE}{\left(1/n\right)\sum_{t=1}^{n} Y_t}\right] \times 100\% \quad , \qquad \text{(S5)}$$

where Y_t is the actual value, y_t is the predicted value, and n is the number of samples.

The values predicted based on the workplace addresses of the study subjects were used as individual long-term air pollution exposure concentrations. For the non-diseased subjects, we assessed exposure levels between the two medical examinations. To reduce the lag effect, the exposure levels of diseased subjects were the values between the last medical examination and the midpoint of the current medical examination.

References

Hoek G, Beelen R, Kos G, Dijkema M, van der Zee S C, Fischer P H, Brunekreef B (2011). Land use regression model for ultrafine particles in Amsterdam. Environmental Science and Technology, 45(2): 622–628

Lee S J, Serre M L, van Donkelaar A, Martin R V, Burnett R T, Jerrett M (2012). Comparison of geostatistical interpolation and remote sensing techniques for estimating long-term exposure to ambient $PM_{2.5}$ concentrations across the continental United States. Environmental Health Perspectives, 120(12): 1727–1732

Appendix C

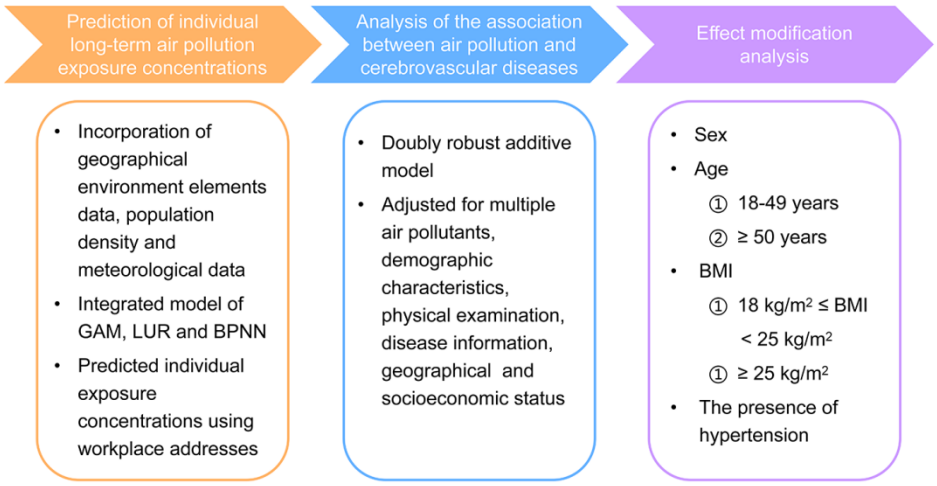

Fig. S1 Schematic diagram of the analysis content and process in this study.