Supplementary Material

Confounding mechanisms and adjustment strategies in air pollution epidemiology: a case study assessment with the UK Biobank cohort

Variable definitions and sources in the UKB database

The data used in the analysis were retrieved from the UKB database. The information was mostly collected during recruitment during the baseline assessment, with the exception of the follow-up and mortality data, which are linked routinely and constantly updated. The other source of data is the time-varying $PM_{2.5}$ exposure that was generated by the spatio-temporal machine learning (ML) model (Ref 12 in the main text), which was linked separately and will be integrated into the UKB database soon. Table S1 below provides information on the selected variables. More details are available in the UKB Showcase (https://biobank.ndph.ox.ac.uk/showcase/), which includes a 'search' tab to gather information on each variable using the related Field IDs.

Table S1. Definition of variables in the UKB database used in the analysis, together with the corresponding Field IDs in the UKB database.

Variable	Field IDs	Notes
Date of birth	34, 52	Derived from year (Field 34) calendar month (Field 52) of birth, and assigned to the 15th day (the exact day is not provided)
Date of death	40000	Acquired from the central registry
Cause of death	40001	Underlying (primary) cause of death, coded using ICD-10
Date lost to follow-up		Date from which a participant is believed to be lost to follow-up (not updated from 2017)
Date of start of follow-up	53	Date when a participant attended a UKB assessment centre
Assessment centre	54	One of the 22 assessment centres at which participants were recruited
Sex	21022	Genetic sex
Age at baseline	31, 22001	Derived based on the date of birth and date of attending the initial assessment centre
Townsend deprivation index	22189	Assigned corresponding to the output area in which the participant's residence postcode is located
Ethnic background	21000	Amalgam of sequential branching questions
Employment	6142, 20119	Field 20119 was used to correct the original Field 6142
Education	6138	Education level based on qualifications
Income	738	Average total household income before tax
Waist-to-hip ratio	48, 49	Ratio of circumferences of waist (Field 48) and hip (Field 49)
Physical activity	22032	(IPAQ score)
Alcohol drinking status	1558	Based on the frequency of alcohol drinking
Smoking status	20116	Current/past smoking status of the participant
Smoking intensity	20161	Pack years of smoking (derived by combining various Fields)
Living alone	709, 670	Derived from information on the number of people in the household (Field 709) and type of accommodation (Field 670)
Urban-rural classification	20118	Derived by linking each participant's home postcode to the 2001 census from the Office of National Statistics

Greenspace	24500	Percentage of land-use type classified as greenspace in a buffer of 1000m around the participant's residence
$PM_{2.5}$ from LUR model	24006	$PM_{2.5}$ level (μg/m³) at each participant's residence in 2010 derived from a Europe-wide land-use regression (LUR) model

Table S2: Hazard ratio (HR) associated with an increase of 10 μg/m³ in $PM_{2.5}$ in the UK Biobank cohort, estimated from models with different adjustment strategies for assessment centre, with 95% confidence intervals (CI).

Assessment centre adjustment	**HR (95% CI)**
Strata (main model)	1.25 (1.06-1.49)
Dummy/indicator variable	1.26 (1.07-1.48)
Normal random effect (default in coxme R package)	1.09 (0.95- 1.26)
Gamma random effect ("frailty" argument in the survival R package)	1.25 (1.04-1.50)

Table S3: Hazard ratio (HR) associated with an increase of 10 μg/m³ in $PM_{2.5}$ (95% confidence intervals) in the UK Biobank cohort, estimated by the full model, the full model without (w/o) lifestyle factors and the full model excluding each lifestyle factor separately.

Adjustment	**HR (95% CI)**
Full adjustment (main model)	1.25 (1.06-1.49)
Main model w/o lifestyle factors	1.34 (1.12-1.59)
Main model w/o smoking status	1.29 (1.08-1.53)
Main model w/o packs cigs/year	1.26 (1.06-1.49)
Main model w/o alcohol intake/week	1.25 (1.05-1.48)
Main model w/o waist-to-hip ratio	1.25 (1.06-1.49)
Main model w/o IPAQ (physical activity)	1.26 (1.06-1.50)
Main model w/o living alone or not	1.27 (1.07-1.51)
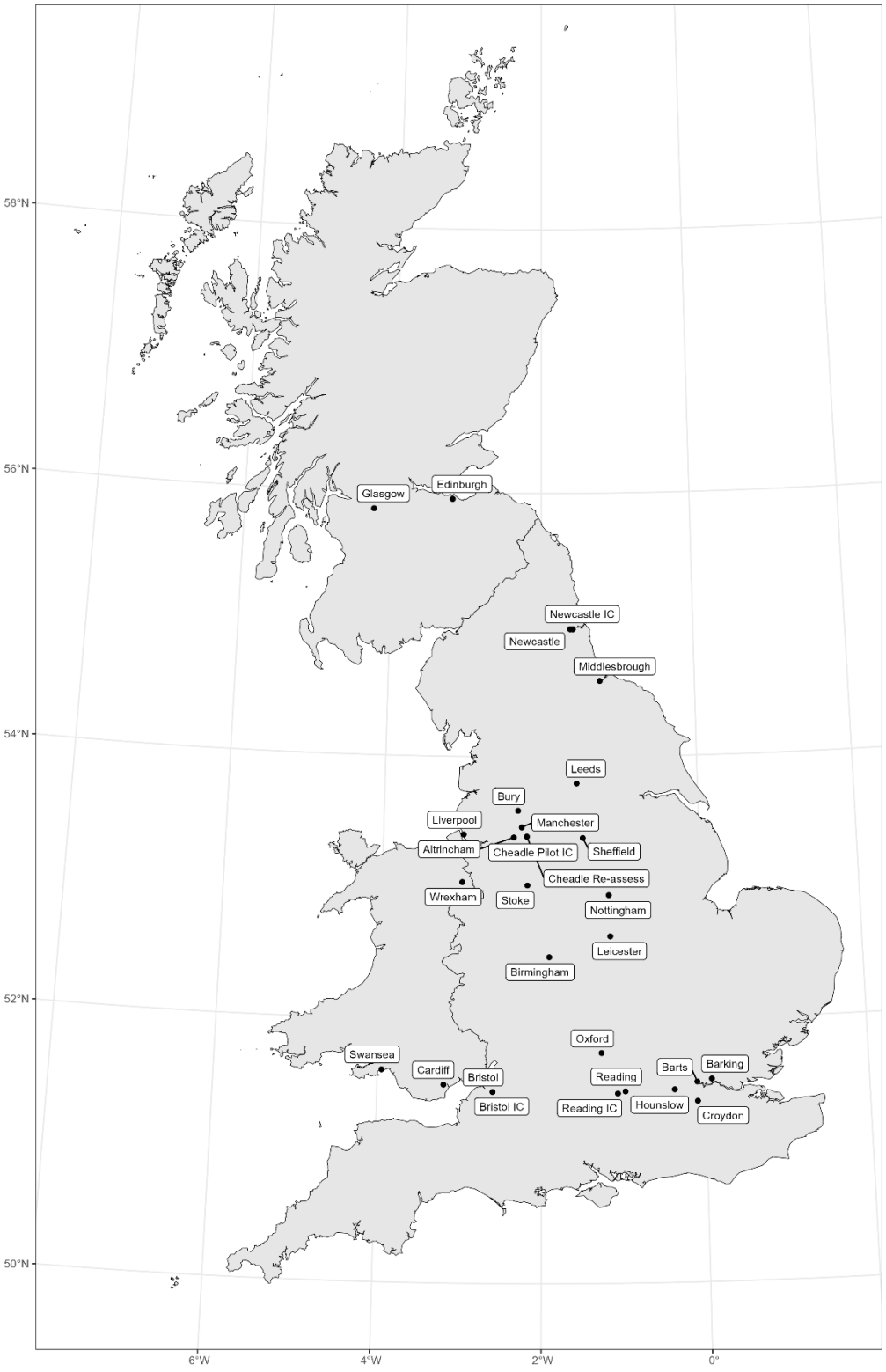
Figure S1: Geographical distribution of the UK Biobank assessment centres.