Supplementary Material

Long-term associations between time-varying exposure to ambient $PM_{2.5}$ and mortality risks: an analysis of the UK Biobank

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 at recruitment during the baseline assessment, with the exception of the follow-up and mortality data that 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 27 in the main text), which was linked separately and will be integrated into the UKB database soon. eTable 1 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.

eTable 1. 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 ($\mu g/m^3$) at each participant's residence in 2010 derived from a Europe-wide land-use regression (LUR) model

Additional Supplementary Tables

eTable 2. Hazard ratios (HRs, with 95% confidence intervals) of different mortality causes associated with an increase of 10 μg/m³ in PM2.5 in the UKB cohort, estimated using different models for modelling the lag-response relationships along an 8-year exposure window (lag 0-7) and the fully-adjusted confounder model (Model 6 in Table 3 in the main text).

	Main model (lag 0-7 average)	Spline-DLM	Strata-DLM
All causes	1.27 (1.06-1.53)	1.29 (1.07-1.55)	1.28 (1.06-1.54)
Non-accidental	1.24 (1.03-1.50)	1.27 (1.05-1.54)	1.26 (1.04-1.52)
Cardiovascular	0.88 (0.59-1.31)	0.92 (0.62-1.38)	0.91 (0.60-1.36)
Respiratory	2.07 (1.04-4.10)	2.28 (1.14-4.57)	2.19 (1.09-4.38)
Lung cancer	1.66 (0.86-3.19)	1.61 (0.83-3.12)	1.64 (0.84-3.17)

Main model: corresponding to a simple DLM with a single stratum modelling a constant lag-response risk

Spline-DLM: using natural cubic splines with 4 degrees of freedom (two-equally-space internal knots and an intercept)

Strata-DLM: using strata defining steps for lags 0, 1-2, 3-4, and 5-7

eTable 3. Comparison of the number of events for different mortality causes and corresponding hazard ratios (HRs, with 95% confidence intervals) associated with an increase of 10 μg/m³ in $PM_{2.5}$, estimated using alternative exposure sources and indices (see Figure 3 in the main text), using the fully-adjusted confounder model (Model 6 in Table 3).

	N	ML time-varying	ML 2010	LUR 2010
All causes	30,750	1.26 (1.04-1.53)	1.13 (0.96-1.33)	1.20 (1.03-1.39)
Non-accidental	28,934	1.22 (1.00-1.49)	1.12 (0.95-1.33)	1.20 (1.03-1.40)
Cardiovascular	6,287	0.87 (0.57-1.34)	0.83 (0.58-1.18)	1.25 (0.90-1.72)
Respiratory	2,262	2.55 (1.24-5.27)	1.58 (0.86-2.92)	1.97 (1.17-3.32)
Lung cancer	2,543	1.57 (0.79-3.11)	1.22 (0.68-2.17)	1.12 (0.68-1.86)

ML-time-varying: annual $PM_{2.5}$ varying during the follow-up period from a spatio-temporal machine learning model (Ref 23 in the main text)

ML-2010: annual $PM_{2.5}$ in 2010 from the model above.

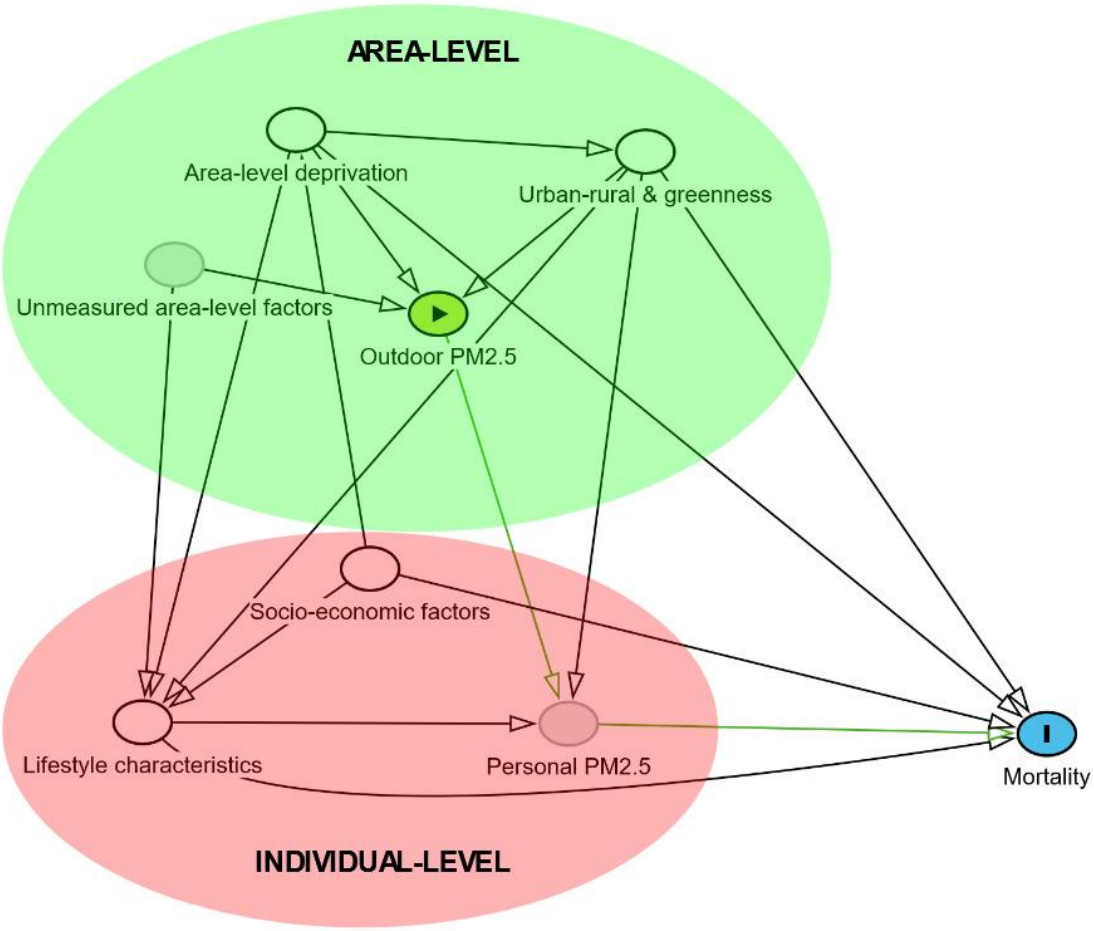
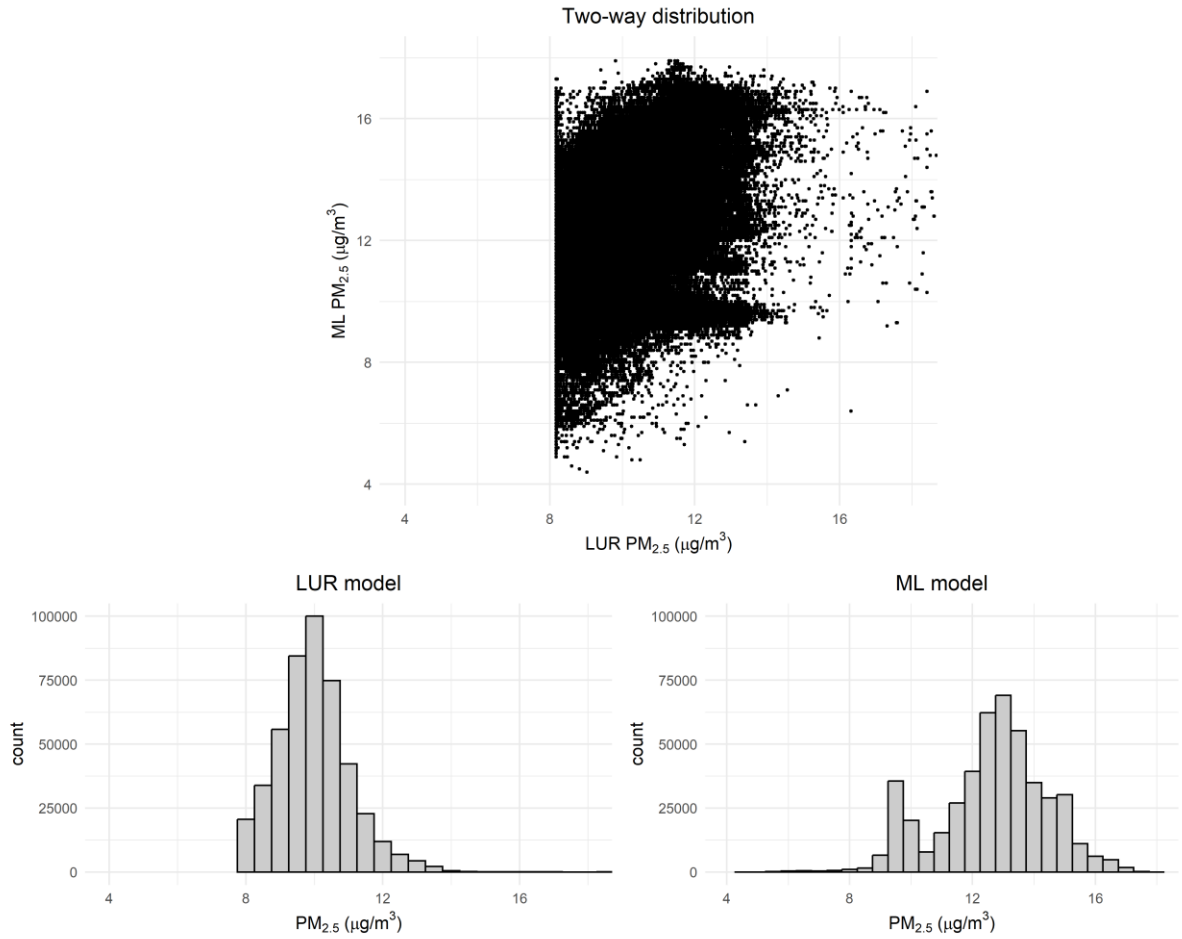
LUR 2010: annual $PM_{2.5}$ in 2010 from a Europe-wide land-use regression (LUR) model (Ref 27 in the main text)

eTable 4. Number of events for different mortality causes and corresponding hazard ratios (HRs, with 95% confidence intervals) associated with an increase of 10 μg/m³ in $PM_{2.5}$, estimated by truncating the follow-up to the pre-2020 (left) and post-2013 (right) periods, using the fully-adjusted confounder model (Model 6 in Table 3 in the main text).

	N	Pre-2020 period	N	Post-2013 period
All causes	29,752	1.28 (1.05-1.56)	26,180	1.33 (1.09-1.62)
Non-accidental	28,238	1.24 (1.02-1.52)	24,423	1.30 (1.06-1.58)
Cardiovascular	6,029	0.85 (0.55-1.30)	5,376	0.99 (0.65-1.52)
Respiratory	2,203	2.22 (1.08-4.56)	2,046	2.38 (1.16-4.89)
Lung cancer	2,578	1.86 (0.94-3.70)	2,002	2.12 (1.07-4.20)

Supplementary Figures

eFigure 1. Directed acyclic graph (DAG) illustrating the assumed causal pathways involving exposure to $PM_{2.5}$ and mortality in the presence of specific sets of other risk factors. The figure distinguishes between area-level and individual-level sets of factors, and it includes both observed (definite) and unobserved (transparent) variables. The graph suggests that all the observed sets of factors (area-level deprivation, urban/rural & greenness, individual-level socio-economic variables, and lifestyle) should be adjusted for in order to avoid potential confounding.



eFigure 2. Comparison of distributions of the annual average exposure to $PM_{2.5}$ in 2010 across participants of the UKB cohort from the spatio-temporal machine learning (ML) model and the Europe-wide land-use regression (LUR) model.