Reconstructing individual-level exposures in cohort analyses of environmental risks: An example with the UK Biobank

Jacopo Vanoli[1,2]✉, Malcolm N. Mistry[2,3], Arturo De La Cruz Libardi[2], Pierre Masselot[2], Rochelle Schneider[2,4], Chris Fook Sheng Ng[5], Lina Madaniyazi[1] and Antonio Gasparrini[2]

[1]School of Tropical Medicine and Global Health, Nagasaki University, Nagasaki, Japan. [2]Environment & Health Modelling (EHM) Lab, Department of Public Health Environments and Society, London School of Hygiene & Tropical Medicine, London, UK. [3]Department of Economics, Ca' Foscari University of Venice, Venice, Italy. [4]Φ-lab, European Space Agency, Frascati, Italy. [5]Department of Global Health Policy, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan. ✉email: jacopo.vanoli@lshtm.ac.uk

Tint Lwin Win

Student | Data Science in Healthcare and Clinical Informatics

18 July 2025

Outlines

- Introduction
- Framework
- Case study
- Discussion
- Limitation

Why link environmental exposure & health?

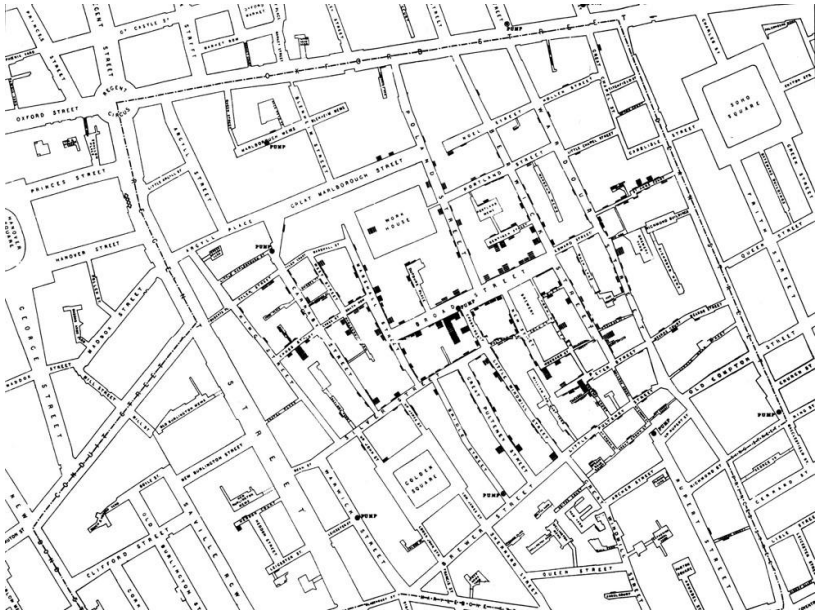
- Environmental exposures are key determinants of health.
- Large cohorts offer power but require robust linkage methods
- Challenge: Accurately assigning accurate exposures at the individual level, time-varying exposure histories in large cohorts.
  - Requires linking high-resolution environmental data with detailed residential histories.
  - Must address issues of data privacy, spatial accuracy, and temporal consistency.
- Examples of exposures: Air pollution, noise, temperature, chemicals.

Objective

- To present a practical, scalable framework for reconstructing individual-level environmental exposures in large cohort studies, using the UK Biobank as a case study.
- Key Questions Addressed:
  - How can we link environmental exposure data to individual participants in a cohort?
  - What are the methodological steps and considerations?
  - How does this approach improve the quality and impact of environmental health research?
- Significance: High-quality exposure assessment is essential for:
  - Understanding the true health risks of environmental factors.
  - Informing effective public health policies and interventions.
  - Advancing the science of environmental epidemiology.

Evolution of environmental epidemiological studies

- Early studies:
  - Early focus on infection diseases: Landmark example: John Snow's 1854 cholera investigation in London, which is widely regarded as the first environmental epidemiology study.
  - Relied on administrative databases and ecological designs.
  - Limited individual-level data, higher risk of misclassification.
- Modern cohorts:
  - High-resolution spatio-temporal exposure maps.
  - Improved linkage procedures and exposure modeling.
  - Large, population-based (e.g., UK Biobank, ESCAPE).
  - Collect detailed lifestyle, genetic, and residential data



Key input data for exposure linkage

- **Cohort baseline information**
  - Enrollment, follow-up, demographics, lifestyle factors
- **Health outcomes**:
  - Hospitalizations, diagnoses, mortality, timing of health events
- **Residential histories**:
  - Detailed addresses with start and end dates of residence
  - Geocoded locations (latitude and longitude)
  - Records of residential mobility throughout the study period
- Environmental exposure data
  - High-resolution spatio-temporal exposure maps for environmental factors.
  - Measured or modelled environmental data linked to geographic locations
  - Data covering the relevant study period with sufficient temporal granularity

Data example – UK Biobank - 1

- Overview
  - Cohort Size: Over 500,000 participants, aged 40–69, recruited between 2006 and 2010 from across the UK.
  - Purpose: To enable large-scale research into genetic, lifestyle, and environmental determinants of health and disease.

Data type	Example content
Baseline info	Enrolment date, follow-up date, demographics
Health outcomes	Hospitalization, diagnoses, ICD-10 codes, event dates
Residential history	Address locations, dates, geocoded coordinates
Environmental exposure map	Annual 1-km grids for $PM_{2.5}$ & NO_2

Data example – UK Biobank - 2

- Residential history data
  - **Collection**: Participants provide address histories, which are geocoded to easting and northing.
  - **Resolution**: Grid coordinates are available at location represents the centroid of a 1km and 100m buffer that contains the exact location to protect privacy.
  - **Updates**: Address changes are recorded through self-report and NHS records.
  - **Privacy**: Exact addresses are not disclosed to researchers; only grid-based locations are used.

Data example – UK Biobank - 3

- Environmental exposure data
  - **Air pollution**: Daily PM2.5 concentrations modeled on a 1x1 km grid for 2008–2018, using machine learning with monitoring stations, satellite data, and land-use information.
  - **Noise**: Annual average noise levels at the residential address.
  - **Other exposures**: Data on green space, traffic proximity, and other local environmental factors

Example of pseudo cohort data, UK Biobank

(a) Cohort info

Subject ID	Enrolment date	Last follow-up date
1	May 1, 2007	March 12, 2017
2	April, 14, 2009	September 25, 2019
3	November 23, 2006	Present

(b) Inpatient visit outcomes table by subject

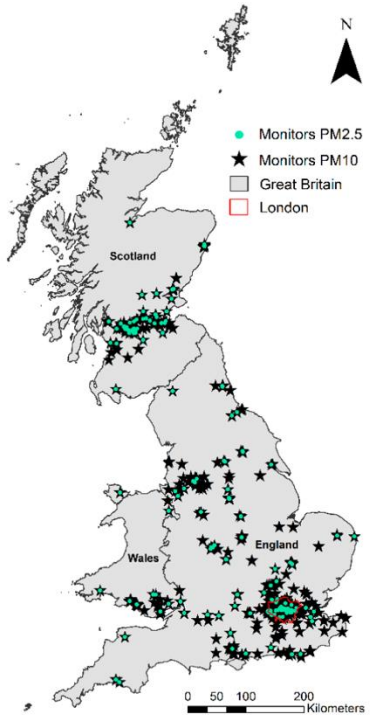
Subject ID	ICD	Date
1	E11	April 23, 2012
1	I20	July 4, 2013
1	I21	September 30, 2016
2	C34	February 24, 2010
3	J40	March 14, 2007
3	J41	April 11, 2008
3	J43	May 22, 2009

(c) Residential histories

Subject ID	Location ID	Start date	End date	Easting	Northing
1	Loc_12	April 1, 2005	May 22, 2012	515,200	184,800
1	Loc_43	May 23, 2012	March 12, 2017	384,800	394,100
2	Loc_92	December 18, 2007	September 3, 2009	342,700	387,100
2	Loc_6	September 4, 2009	April 3, 2017	528,100	105,600
2	Loc_24	April 4, 2017	September 25, 2019	459,900	450,700
3	Loc_87	November 20, 1994	Present	177,500	314,500

Exposure measurement - 1

- **Multi-stage satellite-based machine learning model**
  - **Stage-1** - augments monitor-PM2.5 series using co-located PM10 measures
  - **Stage-2** - imputes missing satellite aerosol optical depth (AOD) observations using atmospheric reanalysis models.
  - **Stage-3 -** integrates the output and spatial and spatio-temporal variables to build a prediction model for PM2.5
  - **Stage-4 -** applies Stage-3 models to estimate daily PM2.5 concentrations over a 1 km grid.



Spatial distribution of 581 PM10 (black star) and 183 PM2.5 (turquoise dots) monitors across Great Britain

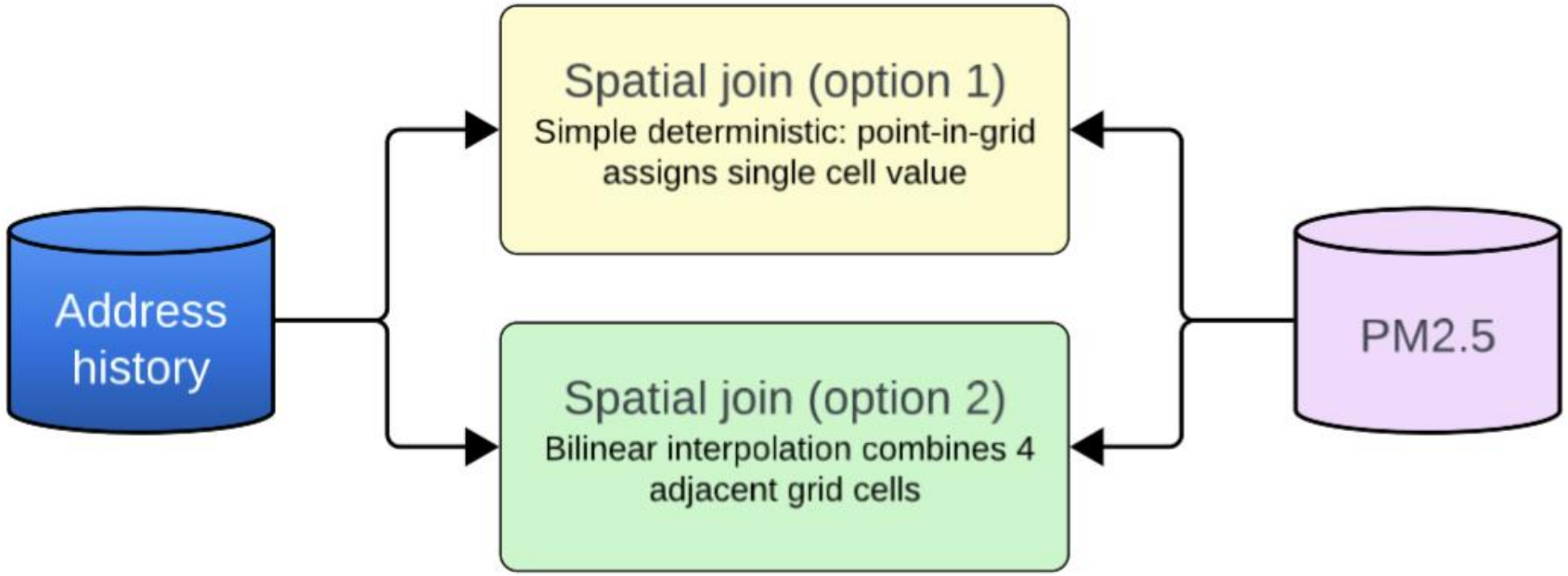
Exposure measurement - 2

- **Predictors:** A long list of spatial and spatio-temporal predictors were used, including:
  - Remote sensing satellite observations
  - Traffic data
  - Weather simulations
  - Road characteristics
  - Land-use information
- **Performance:** The model demonstrated good overall performance with a cross-validated R^2 of 0.767.

Overall data linkage framework

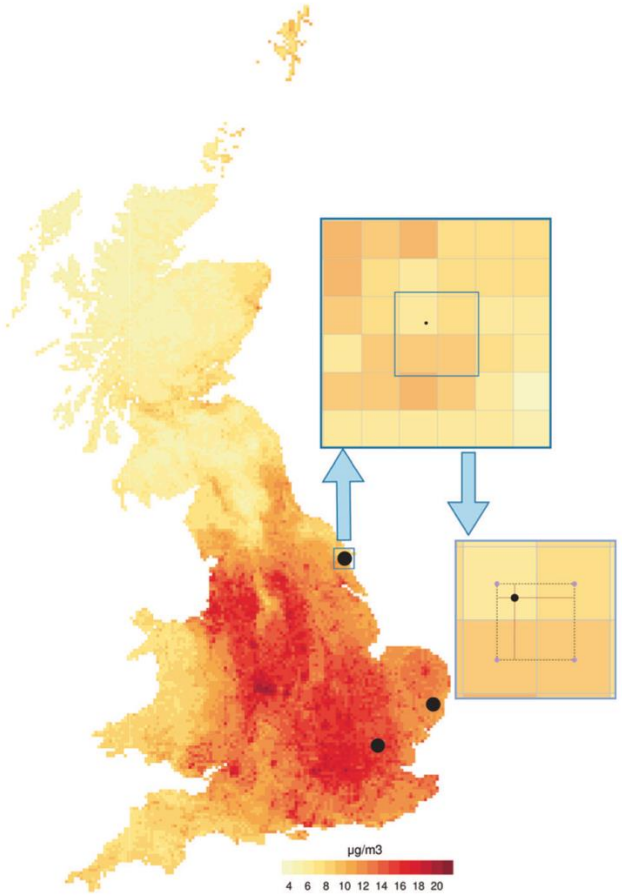
- Deterministic spatial join (address ↔ pollution grid)
- Temporal alignment to generate exposure time-series
- Aggregate exposures for epidemiologic analyses

Step 1: Spatial linkage



Step 1 – Spatial linkage 1

- Address Mapping:
  - Each participant's residential address is geocoded to precise latitude and longitude coordinates.
  - These coordinates are overlaid onto a high-resolution exposure grid 1x1 km for air pollution data.
- Grid cell assignment:
  - The exposure grid consists of regularly spaced centroids, each representing the center of a grid cell with a modeled PM2.5 value
  - For each address, the four nearest grid centroids are identified.



Step 1 – Spatial linkage 2

- Bilinear Interpolation:
  - Bilinear interpolation combines values from four nearest grid points.
- Advantages:
  - **Accuracy**: Incorporates information from multiple grid cells, improving spatial precision.
  - **Privacy**: Prevents back-tracing of exact addresses.

Step 2 – Reconstructing exposure histories

- Integration of residential histories and exposure data:
  - Residential histories (dates, locations).
  - Daily exposure data for each address.
- Accounting for residential mobility:
  - When a participant moves, their exposure profile is updated to reflect the new location.
  - Changes in exposure due to residential moves are accurately captured over time
- Result:
  - Detailed time series of daily exposures for individual, covering the entire follow-up period of the cohort study.

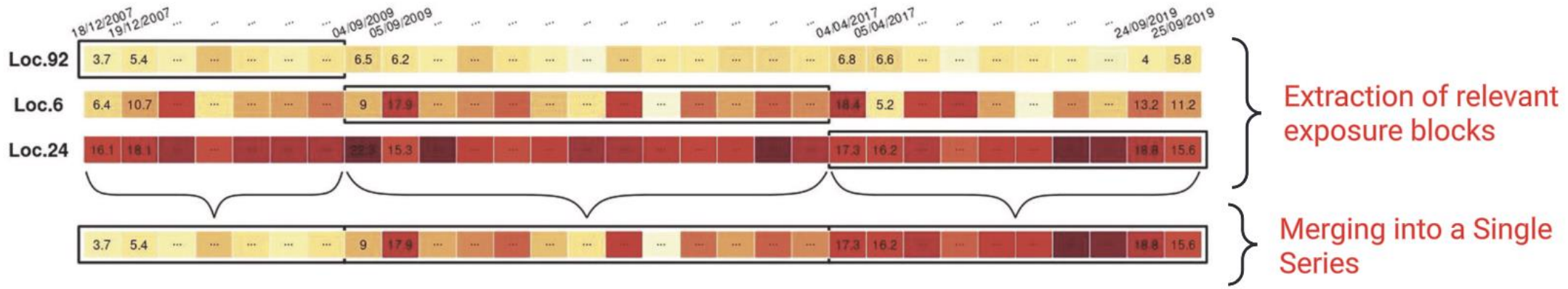
Step 2 - example – exposure series construction

- For each participant:
  - Extract exposure values for each residential period.
  - Concatenate to form a continuous exposure profile over time.
- Visualization: Timeline of exposures reflecting address changes.

Timeline at each location for Subject 2

Address: **Loc.92** 1) 18/12/2007 → 03/09/2009 **Loc.6** 2), 04/09/2009 → 03/04/2017 **Loc.24** 3), 04/04/2017 → 25/09/2019

Daily PM2.5 exposure series location

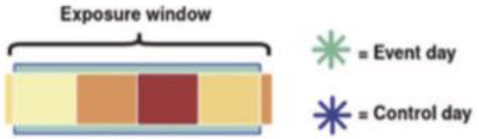


Step 3 – Defining exposure summaries - 1

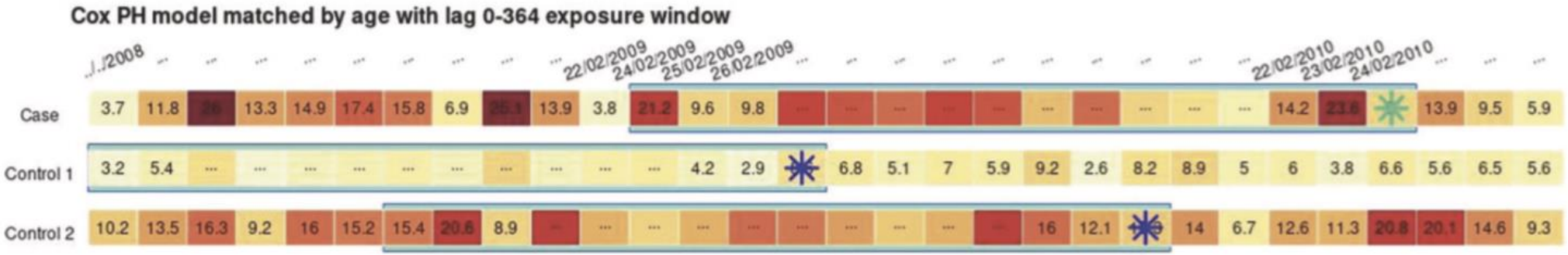
- **Purpose**:
  - To translate detailed, daily individual exposure histories into summary metrics suitable for epidemiological analysis.
- **Long-term exposure summaries:**
  - Capture cumulative or average exposure over extended periods (e.g., 1-year, 5-year, or entire follow-up)
  - Used in studies of chronic effects, such as cardiovascular or cancer risk.
  - Common metrics: Annual average concentration
- **Short-term exposure summaries:**
  - Focus on recent or acute exposures (e.g., same day, previous 1–7 days)
  - Used in studies of acute effects, such as triggering of asthma attacks or myocardial infarction.
  - Common metrics: Lagged daily exposures, moving averages over short windows (e.g., 3-day or 7-day average), maximum daily exposure within a short window

Step 3 – Defining exposure summaries - 2

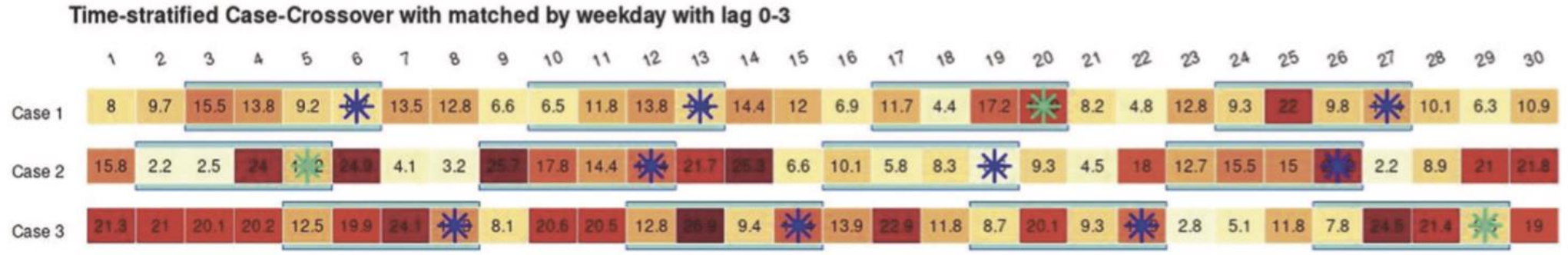
- **Biological relevance:**
  - The chosen summary should match the hypothesized induction or latency period for the health outcome (e.g., years for chronic disease, days for acute events).
- **Study design considerations:**
  - Cohort studies often use long-term summaries for chronic outcomes.
  - Case-crossover or time-series studies use short-term, lagged summaries for acute outcomes.



Example 1: Retrieve exposure history backwards with a lag period, 365-day (lag 0-364) averages, for cancer cases and matched controls



Example 2: Retrieve exposure history backwards with a lag period, 4-day (lag 0-3) averages



PM2.5 and cardiovascular hospital admissions

- **Study population**: analyzed 377,736 adult participants (aged > 40) from the UK Biobank cohort.
- **Exposure Assessment**:
  - Individual level of PM2.5 predictions from a 1-km grid across the UK based on residential history data.
  - Time-varying averages of PM2.5 over 1-year, 3-year, and 5-year windows were investigated.
- **Outcome**: Incidence of major adverse cardiovascular events (MACE), myocardial infarction (MI), heart failure, atrial fibrillation and flutter and cardiac arrest.

Vanoli, J., Quint, J. K., Rajagopalan, S., Stafoggia, M., Al-Kindi, S., Mistry, M. N., Masselot, P., de la Cruz Libardi, A., Fook Sheng Ng, C., Madaniyazi, L., & Gasparrini, A. (2024). Association between long-term exposure to low ambient PM2.5 and cardiovascular hospital admissions: A UK Biobank study. *Environment International*, *192*. https://doi.org/10.1016/j.envint.2024.109011

PM2.5 and cardiovascular hospital admissions (cont.)

- **Statistical Analysis**: Cox proportional hazard model with time-varying predictors.
- **Key Findings:** Positive associations between long-term PM2.5 exposure and several cardiovascular outcomes
  - Hazard Ratio [HR] for 5 µg/m3 increase in PM2.5 - for 5-point MACE of [1.12 (95 %CI: 1.00–1.26)], heart failure [1.22 (1.00–1.50)] and cardiac arrest [1.16 (1.03–1.31)]
  - Could not find any association with acute MI

Vanoli, J., Quint, J. K., Rajagopalan, S., Stafoggia, M., Al-Kindi, S., Mistry, M. N., Masselot, P., de la Cruz Libardi, A., Fook Sheng Ng, C., Madaniyazi, L., & Gasparrini, A. (2024). Association between long-term exposure to low ambient PM2.5 and cardiovascular hospital admissions: A UK Biobank study. *Environment International*, *192*. https://doi.org/10.1016/j.envint.2024.109011

PM2.5 and Mortality

- **Study population**: analyzed 498,090 participants from the UK Biobank cohort.
- **Exposure Assessment**:
  - individual level of PM2.5 predictions from a 1-km grid across the UK based on residential history data.
  - Aggregated into annual series for an 8-year lag window
- **Outcome**: All-cause mortality, nonaccidental mortality, cardiovascular mortality, respiratory mortality, lung cancer mortality

Vanoli, J., de La Cruz Libardi, A., Sera, F., Stafoggia, M., Masselot, P., Mistry, M. N., Rajagopalan, S., Quint, J. K., Ng, C. F. S., Madaniyazi, L., & Gasparrini, A. (2025). Long-term Associations Between Time-varying Exposure to Ambient PM2.5 and Mortality: An Analysis of the UK Biobank. *Epidemiology*, *36*(1), 1–10. https://doi.org/10.1097/EDE.0000000000001796

PM2.5 and Mortality

- An increase of 10 μg/m³ in PM2.5 was associated
  - HR - 1.27 (95% confidence interval:1.06, 1.53) for all-cause mortality
  - HR - 1.24 (1.03, 1.50) for nonaccidental mortality
  - HR - 2.07(1.04, 4.10) for respiratory mortality
  - HR - 1.66 (0.86, 3.19) for lung cancer mortality

Vanoli, J., de La Cruz Libardi, A., Sera, F., Stafoggia, M., Masselot, P., Mistry, M. N., Rajagopalan, S., Quint, J. K., Ng, C. F. S., Madaniyazi, L., & Gasparrini, A. (2025). Long-term Associations Between Time-varying Exposure to Ambient PM2.5 and Mortality: An Analysis of the UK Biobank. *Epidemiology*, *36*(1), 1–10. https://doi.org/10.1097/EDE.0000000000001796

Discussion - 1

- Privacy protection
  - Bilinear interpolation prevents reverse-engineering addresses and minimize the risk of re-identification.
  - Spatial data masking - Use of grid-based or aggregated spatial data (e.g., 1x1 km grids) instead of exact addresses
  - Data de-identification and anonymization
    - Remove or mask direct identifiers (names, addresses, dates of birth).
  - UK Biobank governance & approvals in place
  - Essential for public data releases and ethical compliance

Discussion - 2

- Factors affecting accuracy
  - High-resolution environmental data improve the precision of exposure assignment.
  - Coarse grids or sparse monitoring networks can lead to exposure misclassification, especially in areas with high spatial variability in pollution.
  - Advanced modeling techniques, such as land use regression and machine learning, have increased spatial accuracy.
  - Complete and precise residential histories are crucial for reconstructing individual exposure profiles.
    - Inaccuracies or gaps in address records, and failure to account for residential mobility, can introduce significant error.
  - Quality of residential history (self-report, NHS records).

Discussion - 3

- What are the main challenges in reconstructing individual exposure histories?
  - Ensuring accuracy of residential history data (e.g., address completeness, date precision).
  - Handling residential mobility and gaps in address records.
  - Dealing with uncertainties in exposure modeling and spatial resolution.
- How do different interpolation methods affect exposure assignment and privacy?
  - Choice of interpolation (e.g., bilinear vs. nearest neighbor) impacts spatial precision and potential misclassification.
  - Interpolation methods like bilinear reduce the risk of re-identification by not directly linking exposures to exact addresses.

Strengths of the framework

- **Individual-level precision:**
  - Detailed residential histories and high-resolution environmental data
  - Exposures to be assigned at the individual level rather than aggregated by area.
- **Flexible exposure summaries:**
  - Adaptable to different exposures and study designs.
  - Enables time-varying Cox, case-crossover, Distributed Lag Non-Linear Model (DLNM) analyses
  - Open, reproducible workflow
- **Methodological strengths**
  - Deterministic linkage → high match accuracy
  - Enables fine temporal and spatial analysis.
  - Reduces exposure misclassification bias
- **Privacy-protection:**
  - Safe for use with sensitive cohort data

Applications and extensions

- **Epidemiological Research**
  - Framework can be adapted for various environmental exposure data such as noise, temperature, green space, chemicals.
  - Enables robust analysis of both short-term and long-term health effects of environmental exposures
- **Future directions**:
  - Cancer registries in low- and middle-income countries
  - Future studies may incorporate data from wearable sensors
  - Remote sensing, development of hyperlocal, real-time exposure maps can enhance spatial and temporal resolution.
  - Support for exposome research: By linking multiple exposures across the life course, the framework supports exposome research.

Limitations -1

- Proxy for personal exposure
  - Outdoor vs. Personal exposure: used outdoor environmental levels at residential locations as a proxy for personal exposure, not considering the indoor or occupational exposures can lead to exposure misclassification.
  - Move towards integrating wearable and mobile data
- Exposure model limitation
  - Imperfect predictions, especially in areas with sparse monitoring.
  - Continuous improvement required for prediction accuracy.

Limitations – 2

- Interpolation method selection
  - Lack of rigorous comparison: the choice of bilinear interpolation for spatial linkage was based on practical criteria, no comparison with kriging interpolation, IDW averaging.
- Residential mobility data accuracy
  - Self-Reports: The linkage procedure relies on residential mobility information, from participants' self-reports and NHS records, which can be error-prone and can lead to exposure misclassification.
    - Supplementing self-reported addresses with administrative data sources can improve accuracy
  - Need for accurate, complete residential histories.


Thank You