International Journal of Population Data Science

Cognitive development Respiratory Tract Illness and Effects of eXposure (CORTEX) project: Combining high spatial resolution pollution measurements with individual level data, a methodological approach.

Rodgers, S[1], Lyons, J[2], Mizen, A[2], Berridge, D[2], Akbari, A[3], Carruthers, D[4], Davies, G[1], Dearden, Institute of Fiscal Studies[5], Doherty, R[6], Lake, I[7], Mavrogianni, A[8], Milojevic, A[9], Strickland, S[4], and Wilkinson, P[9]

[1]Swansea University Medical School

[2]Farr Institute, Swansea University Medical School

[3]Health Data Research UK - Wales and Northern Ireland, Swansea University Medical School

[4]Cambridge Environmental Research Consultants

[5]Institute of Fiscal Studies

[6]University of Edinburgh

[7]University of East Anglia

[8]University College London

[9]London School of Hygiene and Tropical Medicine

Introduction

The Secure Anonymised Information Linkage (SAIL) databank facilitated linkage of routinely collected health and education data, high spatial resolution pollution modelling and daily pollen measurements for 18,241 pupils in 7 cross-sectional cohorts across Cardiff city, UK, to investigate effects of air quality and respiratory health conditions on education attainment.

Objectives and Approach

An urban atmospheric dispersion and chemistry modelling system (ADMS-Urban) simulated modelled hourly concentrations of air pollutants: PM2.5, PM10, NO2 and ozone levels. These were summarised into minimum, average and maximum daily readings for 4 time periods (e.g. school hours 9am-3pm) for all home and school locations across Cardiff between 2009 and 2015. The combination of different pollutants, measurements and time-periods created a comprehensive multi-row dataset per location. We transformed the dimensionality of this high-resolution data to create one row of summarised data per pupil per cohort, in preparation for statistical analysis.

Results

157,361 school and home locations across Cardiff were anonymised and household linkage fields were appended to combine pollution estimates at the household/school to individual health data. The pollution dataset contained 369 columns, 472,083 rows per year with one column per location, pollutant type, pollutant measurement, daily time-period, and day of year. Dataset transformation reduced algorithm computation by creating a single date column, producing a five column, 3,446,205,900-row matrix per year dataset. The algorithm adjusted for weekends, school/bank holidays and allowed location to vary 3pm-5pm on school days when pupil location was uncertain. The algorithm calculated tailored pollution exposures per pupil for revision and examination periods, creating one row per pupil and reducing 7 years of data and 24 billion rows to 18,241.

Conclusion/Implications

We successfully linked 95% of the cohorts' household/school pollution data to their corresponding health and education data. This demonstrates data linking retrospective exposures for total populations using multiple daily locations, and extends our analysis platform for natural experiments to include daily exposure. Future work includes adding modelled route exposures.