ABSTRACT

**Background:** Recent epidemiological studies of air pollution have adopted spatially-resolved prediction models to estimate air pollution concentrations at people’s homes. However, the benefit of these models were limited in many studies that used existing health data relying on incomplete addresses resulting from confidentiality concerns or lack of interest when designed. This simulation study aimed to understand the impact of incomplete addresses on health effect estimation based on the association between particulate matter with diameter ≤ 10 µm (PM10) and low birth weight (LBW).

**Methods:** We generated true annual-average concentrations of PM10 at 46,007 mothers’ homes and their LBW status, using the parameters obtained from our data analysis and a previous study in Seoul, Korea. Then, we hypothesized that mothers’ address information is limited to the district and compared the properties of their health effect estimates of PM10 with those using complete addresses. We performed this comparison across eight environmental scenarios that represent various spatial distribution of PM10 and nine exposure prediction methods that provide different sets of predicted PM10 concentrations of mothers.

**Results:** We observed increased bias and root mean square error consistently across all environmental scenarios and prediction methods using incomplete addresses compared to complete addresses. However, the bias related to incomplete addresses decreased when we used population-representative exposures averaged at the district from predicted PM10 at census tract centroids.

**Conclusions:** Our simulation study suggested that individual exposure estimated by prediction approaches and averaged across population-representative points can provide improved accuracy in health effect estimates when complete address data are unavailable.

INTRODUCTION

Long-term exposure to air pollution was associated with mortality and morbidity in many epidemiological studies and the investigation was expanded to large health data such as multi-city or multi-country cohorts1–4. Accurate assessment of individual exposure to long-term air pollution has been crucial in these studies, as individual air pollution measurements are not available given financial and technical constraints. Recent studies developed exposure prediction models to represent substantial spatial variability of exposures across study participants and enhanced the capacity to assess the association with human health. These models produced air pollution concentrations estimated at people’s homes or workplaces as their individual-level exposures5,6. Specifically, physicochemical models relied on emissions and meteorology data and estimated air pollution concentrations on the grid7,8. Statistical models were mostly constructed in pointwise regression including geographical and/or meteorological characteristics, named land use regression9–11. Additional spatial correlation structure was modelled by using geostatistical techniques such as kriging12,13 and spatial smoothing as applied in generalized additive model14.

The benefit of these advanced exposure prediction approaches could be limited when complete address data are unavailable. It is common that address information of exising health data is available at the coarse spatial scale such as census tract, zipcode area, and district1,15–26. Existing cohorts were often not designed to collect full address data1,19. This limitation is more common in administrative health data constructed based on census or public health insurance. Despite their strength of large representative populations that allow examining the association at the national or regional scale, address data were restricted given the concerns of confidentiality 15,16,20–26. For instance, studies using the U.S. Medicare cohort, the Canadian Census cohort, and the Taiwan National Health Insurance Database cohort assessed individual exposure to air pollution at the zipcode or postal code area which is the finest spatial resolution of available address data17,18,21. This incomplete address data may increase exposure misclassification and affect the accuracy and/or precision of health effect estimates.

This simulation study aimed to understand the impact of the incomplete address information on exposure prediction and health effect estimation. In order to achieve applicability and generalizability of the simulation, we designed our study based on a previous epidemiological study of long-term exposure to particulate matter less than or equal to 10 micrometers per diameter (PM10) and low birth weight (LBW) in Seoul, Korea27.

METHODS

Our simulation procedure consists of four steps (Figure S1): 1) exploratory data analyses to obtain parameters for the underlying distributions of PM10 and LBW; 2) generation of true PM10 exposure and LBW status; 3) application of incomplete addresses and estimation of mothers’ exposures; and 4) health effect estimation of LBW for PM10 and comparison of the performance of health effect estimates by complete and incomplete addresses. Following sub-sections provide detailed information of each step. Further details including formulas are provided in the Supplementary Information.

Data Analysis and Parameter Acquisition

We obtained parameters to be used for generating exposure and outcome from the exploratory analysis of air quality regulatory monitoring data for PM10, geographic variables, and birth certificate data in Seoul, Korea, during 2010.28,29 Using the annual average concentrations of PM10, we fitted empirical variogram models and estimated mean and variance parameters. Mean parameters were regression coefficients for five geographic variables that were highly associated with particulate matter in Seoul30. Three variance parameters include range, partial sill, and nugget that indicate the distance in which spatial correlation exists, spatial variability, and non-spatial variability, respectively.30–32 For LBW, we obtained birth certificate data from the Statistical Geographic Information Service operated by the Statistics Korea and computed the proportion of LBW cases to the total births.27 To focus on the spatial variation, we restricted our study period to a single year in 2010 and selected 46,007 mothers who had births in 2010.

Generation of True Exposure and Outcome

We generated the locations of mothers’ homes based on the spatial distribution of the number of births in Seoul (median area and average population in 2010: 21.59 km2 and 412,520), the Capital of South Korea, which is composed of 25 districts, 422 neighbourhoods, and 16,230 census tracts. Because mothers’ addresses in birth certificate data are available at the district level, we treated census tract centroids as mothers’ potential home addresses. We, then, randomly sampled the same number of centroids to those of mothers in each district with the weight of live births across neighbourhoods in each district. These locations were fixed over the simulation.

Assuming that exposure to PM10 follows a Gaussian random field with spatial dependency, we generated true annual-average PM10 concentrations using mean and variance parameters at all locations (Table 1). These locations include 46,007 mothers’ homes, 37 air quality regulatory monitoring sites, 25 district governmental offices, 422 neighbourhood community centers, 16,230 census tract centroids, and 610 centroids on the 1-km grid in Seoul.

To represent possibly different spatial structures of true PM10 annual-average concentrations, we used different combinations of mean and variance paramters and constructed eight environmental scenarios (ES1-ES8). Eight combinations of paramters gave varying contributions of the mean structure, spatial variability, and non-spatial variability of PM10 to total variability (Table 1, Figure S1). While ES1-ES4 were defined based on different spatial correlation structures, ES5-ES8 additionally include different mean structures characterized by five geographic variables that were highly associated with particulate matter in Seoul30. ES8 was contructed by the optimal paramters from our data analysis.

For outcome, we assumed an inverse logit function as the underlying distribution of LBW. Then, we generated LBW status of mothers using simulated true PM10 concentrations, the proportion of LBW cases, and the effect estimate of LBW for PM10 obtained from our previous study27.

Exposure Prediction

Using simulated PM10 at 37 regulatory monitoring sites, we applied nine prediction methods to estimate mother’s individual exposure to PM10 by complete and incomplete addresse conditions. When mothers’ complete home addresses are available, we applied four prediction methods commonly used in previous studies6,10,11. In the nearest monitor (NM) and inverse distance weighted average (IDWA) methods soly based on (“simulated”) measurements, we assigned PM10 at the monitoring site nearest to a home of each mother and averaged across the sites weighted by inverse squared Euclidean distance from each home, respectively. The other two approaches employed modelling approaches including geographic characteristics that represent direct or indirect pollution sources. Land use regression (LUR) includes these characteristics as predictors in regression equations. Universal kriging (UK) additionally includes spatial correlation as a geostatistical method which optimally derives interpolated concentrations based on mean structure and spatial correlation. We built LUR and UK models using the same five geographic variables to those used in the generation of true PM10 (See the “Data Analysis and Parameter Acquisition” section). Out of the 37 regulatory monitoring sites in Seoul, we used underlying PM10 concentrations from 25 urban-background sites for NM, IDWA, and area averaging (AA), and from all 37 sites including 12 urban roadside sites for LUR and UK. We predicted PM10 concentrations of mothers using estimated regression and/or variance parameters in LUR and UK along with geographic variables at mothers’ homes.

When address data were assumed to be incomplete and available at the district level, we applied one measurement-based and four model-based prediction approaches. As a measurement-based approach, AA assigned the average concentration across all monitoring sites in a district to all mothers living in the same area33,34. Since the regulatory monitoring network in Seoul had one urban-background site in every district, we treated the PM10 concentration at a single site as a special case of AA. In addition, we applied UK to compute area-level representaitve exposure and developed four approaches. Here, we assumed when a pointwise prediction model is available but complete address data are unavailable, a preferred option could be the aggregation of predictions at many representative points35,36. We used three representative locations for aggregation: 422 neighbourhood community centers (UKNA), 16,230 census tract centroids (UKCA), and 610 1-km grid coordinates (UKGA). We predicted PM10 concentrations using UK at these three types of locations, computed district averages, and assigned to the mothers living in the same districts. UKNA and UKCA predictions represent population exposure at the fine spatial scale, whereas UKGA predictions focus on spatially-representative exposure based on spatially even distribution of PM10. We also used predictions at 25 district governmental offices without aggregation (UKD) for comparison.

Health Effect Estimation and Comparison of Properties

Using true and predicted PM10 as well as true LBW status of mothers, we estimated the health effects of LBW for PM10 using logistic regression. Then, we repeated the whole procedure from exposure generation to health effect estimation 1,000 times, and computed properties of health effect estimates over 1,000 simulations as bias, root mean square error (RMSE), average standard error (ASE), coverage probability (CP), and true positive rate (TPR). TPR was computed as the ratio of the number of simulations that provide significantly positive health effect estimates for each predicted PM10 (p-value < 0.05) to that for true PM10. While bias, RMSE, ASE, and CP aim to evaluate the accuracy or uncertainty of the estimates, TPR can indicate statistical power. Finally, we compared the health effect estimate properties between complete and incomplete addresses depending on the exposure prediction method and pollution environment.

RESULTS

True and Predicted PM10

Table S1 and Figure 1 summarize true and predicted annual-average PM10 concentrations at home addresses of 46,007 mothers by different ESs and exposure prediction methods. Mothers’ PM10 concentrations predicted at their homes (mean= 47.25-60.52 µg/m3, standard deviation (SD)= 1.04-6.45 µg/m3) were generally similar on average but less variable compared to true concentrations (46.90-58.55 µg/m3, 4.66-6.45 µg/m3). Variability was even smaller when address information was restricted to the district (SD= 1.04-4.56 µg/m3), compared to complete addresses (1.56-6.45 µg/m3). This pattern was similar across eight ESs with the smallest mean and variability in ES4 where there is no spatial variability in exposure. Across nine prediction methods, the correlation with true expsoure was generally higher in UK and UKCA (Pearson correlation coefficient= 0.26-0.70), with complete and incomplete address data respectively, compared to the other prediction methods (0.00-0.65) across all ESs except ES4 that showed little correlation (Table S2, Figure S3). NM, IDWA, AA, and UKD gave relatively high correlation and slopes close to 1 when there was no mean structure with some spatial correlation in ES1-ES3, but low correlation otherwise. In contrast, LUR provided high correlation (0.60-0.65) when there was a dominant mean structure as shown in ES7 and ES8, but low correlations less than 0.1 without a mean structure.

Health Effect Estimate Properties by Address Availability

Performance of effect estimates of LBW for true and predicted annual-average PM10 concentrations became worse when data availability for PM10 or address was limited. Table 2 shows bias, RMSE, ASW, and CP of health effect estimates in four ESs including ES2, ES3, ES5, and ES8 where different exposure environments are more distinct: Tables S3 and S4, and Figures S4 and S5 show all eight scenarios. Bias and RMSE tended to increase using predicted exposure compared to true exposure, while there was slightly larger increase with incomplete addresses than with complete addresses. Performance varied more across different prediction methods and environmental scenarios under complete addresses than incomplete addresses. Regardless of address availability, CPs were close to 0.95. TPR was generally lower with incomplete addresses than complete addresses (Figure S6).

Health Effect Estimate Properties by Exposure Prediction

All prediction methods under complete addresses, and AA and UKD under incomplete addresses mostly showed negative bias indicating under-estimated health effects (Tables 2, S3, and S4). Bias was particularly large for NM, AA, and UKD that relied on the measurement or prediction at a single monitoring site to assess an individual exposure. However, UKD using predictions estimated at governmental offices that were largely located in highly populated areas gave smaller bias than NM and AA based on regulatory monitoring sites. When address data were fully available, UK gave lower bias and RMSE than other approaches across all ESs. This good performance was notably prominent in ES5 to ES8 which includes a mean structure. NM tended to provide small RMSE and ASE, but relatively large negative bias, while IDWA gave relatively small bias but large RMSE and ASE. LUR provided good performance only when there is a mean structure in the true environment (ES5 to ES8). When address data were limited to the district, three UK-based district averages showed much smaller bias with either direction and slightly higher RMSE and ASE compared to the other two prediction methods of AA and UKD. Among these three approaches, UKCA as the exposure averaged over a large number of population-representative points at the fine spatial scale showed better performance than UKNA and UKGA based on coarse spatial-scale population-represnetative points and spatially-representative points, respectively. TPR was also generally higher in UK and UK-based district averages (Figure S6).

Health Effect Estimate Properties by Environmental Scenarios

Better performance of UK and UKCA under complete and incomplete address conditions, respectively, were consistent across all eight ESs (Tables S3 and S4, and Figure S4-S6). Large bias in NM, AA, and UKD, and large uncertainty in UKNA and UKGA, possibly resulting from small variability in PM10, were also consistent across all ESs. ES4 showed large bias and RMSE and small TPR across all prediction methods.

DISCUSSION

This study focused on the impact of limited availability of address data on health effect estimation compared to complete availability. After hypothesizing that address data availability affects health effect analysis of predicted exposure, we explored the impact of address availability on the performance of health effect estimates depending on exposure prediction methods and different environmental scenarios based on the real-world example of the association between PM10 and LBW. Eight environmental scenarios represented various pollution environments related to different contribution of geographic features and spatial dependency. Furthermore, nine prediction methods exhibited commonly-applied approaches of individual exposure assessment given limited monitoring data with and without additional limitation in address data. Our study showed that the UK-based method can be a good option in studies using limited address data in order to achieve a similar level of accuracy in health effect analysis to that using complete address information

Our simulation study intended to answer an important question that can help inference of epidemiological studies of air pollution relying on limited address data of subjects. Even though many recent studies developed advanced exposure prediction models and allowed the estimation of air pollution concentrations at people’s homes or work places, the benefit of this advance could be limited in many epidemiological studies that are based on existing cohorts and/or administrative health data with incomplete address information. As recent epidemiological studies of air pollution expanded their spatial and temporal coverage to the national or regional scale and to the past several decades, the reliance on existing health data lacking complete address information has become even greater. However, there have been few studies that investigated the impact of limited address information on health analysis and provided realistic guidance. For example, recent two nationwide U.S. cohort studies including limited address data applied prediction models to estimate individual-level long-term PM2.5 concentrations at census tract- and zipcode-level addresses of Medicare cohort participants, and reported the association with total mortality37. Our simulation findings of negative bias using single points in administrative areas suggest the possibility of underestimated health effects in such studies.

Our findings generally showed that kriging-based approaches gave good performance in health effect estimates consistently across different air pollution environments, when individual air pollution measurements are not available. While UK showed better performance compared to other prediction approaches when complete address data are available, UK averaging approaches outperformed with individual address data limited to the district. A possible explanation is that UK modelled by using both mean and variance structures well characterizes air pollution conditions at people’s residences even when there is no mean structure38. In addition, employment of population-representative locations and the following averaging process under the unavailability of precise residential addresses possibly minimized the impact of exposure misclassification. Bias was the smallest and also non-systematic as opposed to other prediction methods that gave consistently negative bias. Out of three UK averaging approaches, UKCA based on UK predictions at census tract centroids gave the lowest RMSE and ASE which were comparable to those of other prediction approaches under the complete address condition. UKCA also showed comparable TPR to those with complete addresses, while it was less likely to detect statistically significant health effect estimates overall with incomplete addresses. However, the benefit of UK-based averaging could be reduced, when we use predictions at the locations including those poorly represented for population as shown in UKGA.

All prediction methods except for UK-based averaging generally showed underestimated health effects given limited exposure or address data. This underestimation can be explained by exposure measurement error derived by poor characterization of individual exposure in prediction models39,40. In our simulation, prediction methods heavily relying on a mean structure such as LUR gave greater underestimation when there is no mean structure in true exposure scenarios, while simple prediction approaches using measurements only shown as NM and IDWA gave larger underestimation when there is a mean structure. Prediction methods using a single location based on the nearest monitor (NM), or district governmental office (UKD) also gave larger negative bias than other methods. Bias was larger in AA and NM based on regulatory monitoring sites than UKD using population-representative locations. In addition, poor assessment of individual exposure can result from poor representativeness of prediction points used for averaging. Our study showed increased positive or negative bias in UKGA using grid coordinates than UKCA based on census tract centroids. Previous simulation studies reported that measurement error derived by a spatial misalignment between regulatory monitoring sites and people’s residences affected misspecification of prediction models and resulted in positive or negative bias in following health effect analysis41–43. Our finding of large bias in UKGA suggested a possible impact of spatial misalignment between prediction points and people’s residences when complete address data are unavailable and prediction models are applied to many representative points.

Our simulation using various environmental scenarios and parameters obtained from data analyses suggests possible generalization of our findings to other pollutants and/or study areas. Although we focused on PM10 which is well known as a regional pollutant with relatively weak mean structure and large spatial correlation, we constructed seven environmental scenarios by assuming different spatial structures in addition to the ES8 based on the parameters estimated directly from the regulatory monitoring data in Seoul. This variation of spatial structure possibly represents more local or regional pollutants with different impacts of local sources or spatial homogeneity, and allows us to apply our findings to other pollutants and/or different regions. In addition, our reliance on real-world data can improve the practical applicability of our simulation findings.

Our study includes several limitations to be further investigated in future research. First, we used low birth weight and logistic regression. Future studies should confirm whether our suggestions are consistent to different health outcomes and health analysis models. Second, we assumed mothers’ residential addresses fixed. Future studies should incorporate mobility and investigate the sensitivity of our findings.

In conclusion, this simulation study suggests that exposure prediction approaches well representing geographic environments and supplemented with population-representative prediction locations can improve accuracy in health effect estimation when complete individual address data are not available. Our findings also provide guidance for a preferred approach to improve the inference in future large-scale epidemiological studies of long-term air pollution.

**References**

1. Beelen, R. *et al.* Effects of long-term exposure to air pollution on natural-cause mortality: an analysis of 22 European cohorts within the multicentre ESCAPE project. *Lancet* **383**, 785–795 (2014).

2. Cesaroni, G. *et al.* Long term exposure to ambient air pollution and incidence of acute coronary events: prospective cohort study and meta-analysis in 11 European cohorts from the ESCAPE Project. *BMJ* **348**, f7412 (2014).

3. Hoek, G. *et al.* Long-term air pollution exposure and cardio- respiratory mortality: a review. *Environ. Heal.* **12**, 43 (2013).

4. Di, Q. *et al.* Air Pollution and Mortality in the Medicare Population. *N. Engl. J. Med.* **376**, 2513–2522 (2017).

5. Hoek, G. *et al.* A review of exposure assessment methods for epidemiological studies of health effects related to industrially contaminated sites. *Epidemiol. Prev.* **42**, 21–36 (2018).

6. Hoek, G. Methods for Assessing Long-Term Exposures to Outdoor Air Pollutants. *Curr. Environ. Heal. Reports* **4**, 450–462 (2017).

7. Binkowski, F. S. & Roselle, S. J. Models-3 Community Multiscale Air Quality (CMAQ) model aerosol component 1. Model description. *J. Geophys. Res. Atmos.* **108**, (2003).

8. Hanha, S. R. Air Quality Model Evaluation and Uncertainty. *JAPCA* **38**, 406–412 (1988).

9. Brauer, M. *et al.* Estimating Long-Term Average Particulate Air Pollution Concentrations: Application of Traffic Indicators and Geographic Information Systems. *Epidemiology* **14**, 228–239 (2003).

10. Hoek, G. *et al.* A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmos. Environ.* **42**, 7561–7578 (2008).

11. Jerrett, M. *et al.* A review and evaluation of intraurban air pollution exposure models. *J. Expo. Sci. Environ. Epidemiol.* **15**, 185–204 (2005).

12. Jerrett, M. *et al.* Spatial Analysis of Air Pollution and Mortality in Los Angeles. *Epidemiology* **16**, 727–736 (2005).

13. Sampson, P. D. *et al.* A regionalized national universal kriging model using Partial Least Squares regression for estimating annual PM2.5 concentrations in epidemiology. *Atmos. Environ.* **75**, 383–392 (2013).

14. Paciorek, C. J., Yanosky, J. D., Puett, R. C., Laden, F. & Suh, H. H. Practical large-scale spatio-temporal modeling of particulate matter concentrations. *Ann. Appl. Stat.* **3**, 370–397 (2009).

15. Seong, S. C. *et al.* Cohort profile: the National Health Insurance Service-National Health Screening Cohort (NHIS-HEALS) in Korea. *BMJ Open* **7**, e016640 (2017).

16. Okayama, A. *et al.* Dietary sodium-to-potassium ratio as a risk factor for stroke, cardiovascular disease and all-cause mortality in Japan: the NIPPON DATA80 cohort study. *BMJ Open* **6**, e011632 (2016).

17. Peters, P. A. *et al.* Data resource profile: 1991 Canadian Census Cohort. *Int. J. Epidemiol.* **42**, 1319–1326 (2013).

18. Jung, C.-R., Lin, Y.-T. & Hwang, B.-F. Ozone, particulate matter, and newly diagnosed Alzheimer’s disease: a population-based cohort study in Taiwan. *J. Alzheimer’s Dis. JAD* **44**, 573–584 (2015).

19. Laden, F., Schwartz, J., Speizer, F. E. & Dockery, D. W. Reduction in Fine Particulate Air Pollution and Mortality. *Am. J. Respir. Crit. Care Med.* **173**, 667–672 (2006).

20. Kim, O. J., Kim, S. Y. & Kim, H. Association between long-term exposure to particulate matter air pollution and mortality in a South Korean national cohort: Comparison across different exposure assessment approaches. *Int. J. Environ. Res. Public Health* **14**, (2017).

21. L., Z. S., Francesca, D., Aidan, M. & M., S. J. Mortality in the Medicare Population and Chronic Exposure to Fine Particulate Air Pollution in Urban Centers (2000–2005). *Environ. Health Perspect.* **116**, 1614–1619 (2008).

22. Carey, I. M. *et al.* Mortality Associations with Long-Term Exposure to Outdoor Air Pollution in a National English Cohort. *Am. J. Respir. Crit. Care Med.* **187**, 1226–1233 (2013).

23. H., F. P. *et al.* Air Pollution and Mortality in Seven Million Adults: The Dutch Environmental Longitudinal Study (DUELS). *Environ. Health Perspect.* **123**, 697–704 (2015).

24. Huss, A., Spoerri, A., Egger, M., Röösli, M. & Group, S. N. C. S. Aircraft noise, air pollution, and mortality from myocardial infarction. *Epidemiology* **21**, 829–836 (2010).

25. Crouse, D. L. *et al.* Ambient PM2.5, O₃, and NO₂ Exposures and Associations with Mortality over 16 Years of Follow-Up in the Canadian Census Health and Environment Cohort (CanCHEC). *Environ. Health Perspect.* **123**, 1180–1186 (2015).

26. Hansell, A. *et al.* Historic air pollution exposure and long-term mortality risks in England and Wales: prospective longitudinal cohort study. *Thorax* **71**, 330–338 (2016).

27. Choe, S.-A., Jang, J., Kim, M. J., Jun, Y.-B. & Kim, S.-Y. Association between ambient particulate matter concentration and fetal growth restriction stratified by maternal employment. *BMC Pregnancy Childbirth* **19**, 246 (2019).

28. National Institute of Environmental Research. 2016 NIER Annual Report. (2017).

29. Yi, S.-J. *et al.* Association between Exposure to Traffic-Related Air Pollution and Prevalence of Allergic Diseases in Children, Seoul, Korea. *BioMed Research International* (2017).

30. Min, K. D., Kwon, H. J., Kim, K. S. & Kim, S. Y. Air pollution monitoring design for epidemiological application in a densely populated city. *Int. J. Environ. Res. Public Health* **14**, 1–12 (2017).

31. Cressie, N. *Statistics for Spatial Data*. (Wiley-Interscience, 2015).

32. Eum, Y., Song, I., Kim, H.-C., Leem, J.-H. & Kim, S.-Y. Computation of geographic variables for air pollution prediction models in South Korea. *Environ. Health Toxicol.* **30**, (2015).

33. Dockery DW, Pope CA 3rd, Xu X, et al. Association between Air Pollution and Mortality in Six U.S. Cities. *N. Engl. J. Med.* **29**, 1230–5 (1993).

34. Pope, C. A. *et al.* Particulate air pollution as a predictor of mortality in a prospective study of U.S. Adults. *Am. J. Respir. Crit. Care Med.* **151**, 669–674 (1995).

35. Pope, C. A. *et al.* Erratum: Mortality risk and fine particulate air pollution in a large, representative cohort of U.S. adults (Environ Health Perspect, (2019), 127, 7, 10.1289/EHP4438). *Environ. Health Perspect.* **127**, 099002–1 (2019).

36. Kim, O. J., Lee, S. H., Kang, S. H. & Kim, S. Y. Incident cardiovascular disease and particulate matter air pollution in South Korea using a population-based and nationwide cohort of 0.2 million adults. *Environ. Heal. A Glob. Access Sci. Source* **19**, 1–12 (2020).

37. Di, Q. *et al.* Air Pollution and Mortality in the Medicare Population. *N. Engl. J. Med.* **376**, 2513–2522 (2017).

38. Kim, S. Y., Sheppard, L. & Kim, H. Health effects of long-term air pollution: Influence of exposure prediction methods. *Epidemiology* **20**, 442–450 (2009).

39. Sheppard, L. *et al.* Confounding and exposure measurement error in air pollution epidemiology. *Air Qual. Atmos. Health* **5**, 203–216 (2012).

40. Szpiro, A. A., Sheppard, L. & Lumley, T. Efficient measurement error correction with spatially misaligned data. *Biostatistics* **12**, 610–623 (2011).

41. Szpiro, A. A., Paciorek, C. J. & Sheppard, L. Does more accurate exposure prediction necessarily improve health effect estimates? *Epidemiology* **22**, 680–685 (2011).

42. Szpiro, A. A. & Paciorek, C. J. Measurement error in two-stage analyses, with application to air pollution epidemiology. *Environmetrics* **24**, 501–517 (2013).

43. Lee, A., Szpiro, A., Kim, S. Y. & Sheppard, L. Impact of preferential sampling on exposure prediction and health effect inference in the context of air pollution epidemiology. *Environmetrics* **26**, 255–267 (2015).

**Figure Legends**

**Figure 1.** Box-plots of true (TE: true exposure) and predicted (NM: nearest monitor, IDWA: inverse distance weight average, LUR: Land-use regression, AA: area average, UK: universal kriging, UKD: UK prediction at governmental offices; UKNA: district average based on UK predictions at 422 neighbourhood community centers; UKCA: district average of UK predictions at 16,230 census tract centroids; UKGA: district average of UK predictions at 610 1-km grid coordinates) annual-average PM10 concentrations at home addresses of 46,007 mothers by eight environmental scenarios (ES1-ES8) in the 1st simulation (blue boxes for true exposure; yellow and red boxes for predicted exposure with complete and incomplete addresses, respectively).

**Table 1.** Spatial characteristics of eight environmental scenarios (ESs) based on their variability components and variance parameters used for simulating true PM10 annual average concentrations

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ES | Variability componenta | | | Variance parameter | | |
| Mean  structure | Spatial variability | Non-spatial variability | Nugget | Partial sill | Range (m) |
| ES1 | None | Dominant | Little | 1.00 | 30.94 | 5,885 |
| ES2 | None | High | Low | 6.86 | 28.98 | 9,609 |
| ES3 | None | Low | High | 11.51 | 34.71 | 20,355 |
| ES4 | None | Little | Dominant | 22.00 | 13.77 | 27,000 |
| ES5 | Moderate | High | Low | 1.00 | 16.98 | 2,524 |
| ES6 | Moderate | Low | High | 6.86 | 12.17 | 4,820 |
| ES7 | Dominant | High | Low | 1.00 | 10.00 | 1,100 |
| ES8 | Dominant | Low | High | 6.86 | 3.60 | 1,004 |

aSpatial characteristics determined by contribution of three variability components (mean structure, and spatial and non-spatial variability) to total variability

**Table 2.** Properties (Bias, RMSE, ASE, and CP) of health effect estimates of true and predicted PM10 concentrations on low birth weight over 1,000 simulations by address availability, exposure prediction methods, and environmental scenarios (ES2, ES3, ES5, and ES8)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | ES2 |  |  |  | ES3 |  |  |  | ES5 |  |  |  | ES8 |  |  |  |
| Address  availability | Exposure  prediction | Biasa | RMSEb | ASEc | CPd | Bias | RMSE | ASE | CP | Bias | RMSE | ASE | CP | Bias | RMSE | ASE | CP |
|  | TE | -0.04 | 1.48 | 0.78 | 0.95 | 0.03 | 1.48 | 0.75 | 0.95 | 0.03 | 1.50 | 0.78 | 0.95 | 0.00 | 1.40 | 0.71 | 0.95 |
| Complete | UK | -0.03 | 2.41 | 1.45 | 0.94 | -0.05 | 2.72 | 1.58 | 0.94 | 0.01 | 2.06 | 1.10 | 0.95 | -0.01 | 1.73 | 0.88 | 0.96 |
| Incomplete | AA  UKD  UKNA  UKCA  UKGA | -0.15  -0.08  -0.01  0.00  0.09 | 1.55  2.42  3.27  2.84  3.83 | 0.83  1.41  1.95  1.74  2.38 | 0.94  0.94  0.94  0.94  0.94 | -0.14  -0.09  0.07  0.05  0.04 | 1.56  2.75  4.08  3.32  4.83 | 0.80  1.62  2.67  2.00  3.63 | 0.95  0.94  0.94  0.95  0.94 | -0.19  -0.16  0.04  0.01  0.05 | 1.76  2.02  3.10  2.33  3.71 | 0.95  1.07  1.68  1.26  2.04 | 0.95  0.94  0.94  0.94  0.95 | -0.25  -0.24  0.06  0.02  -0.11 | 2.04  1.93  4.07  2.23  3.31 | 1.05  1.03  2.14  1.16  1.73 | 0.95  0.95  0.94  0.96  0.95 |

a Bias multiplied by 100

b Root mean square error (RMSE) multiplied by 100

c Average standard error (ASE) multiplied by 100

d Coverage probability (CP) of 95% confidence interval

e TE: true exposure; UK: universal kriging; AA: area-average UKD: UK prediction at governmental offices without aggregation; UKNA: district average based on UK predictions at 422 neighbourhood community centers; UKCA: district average of UK predictions at 16,230 census tract centroids; UKGA: district average of UK predictions at 610 1-km grid coordinates**Figure 1.**

|  |  |  |  |
| --- | --- | --- | --- |
| ES 1 | ES 2 | ES 3 | ES 4 |
|  |  |  |  |
| ES 5 | ES 6 | ES 7 | ES 8 |
|  |  |  |  |