


RESEARCH

Open Access

# Assessing additive effects of air pollutants on mortality rate in Massachusetts



Yaguang Wei<sup>1\*</sup> , Brent Coull<sup>2</sup>, Petros Koutrakis<sup>1</sup>, Jiabei Yang<sup>3</sup>, Longxiang Li<sup>1</sup>, Antonella Zanobetti<sup>1</sup> and Joel Schwartz<sup>1,4</sup>

## Abstract

**Background:** We previously found additive effects of long- and short-term exposures to fine particulate matter (PM<sub>2.5</sub>), ozone (O<sub>3</sub>), and nitrogen dioxide (NO<sub>2</sub>) on all-cause mortality rate using a generalized propensity score (GPS) adjustment approach. The study addressed an important question of how many early deaths were caused by each exposure. However, the study was computationally expensive, did not capture possible interactions and high-order nonlinearities, and omitted potential confounders.

**Methods:** We proposed two new methods and reconducted the analysis using the same cohort of Medicare beneficiaries in Massachusetts during 2000–2012, which consisted of 1.5 million individuals with 3.8 billion person-days of follow-up. The first method, weighted least squares (WLS), leveraged large volume of data by aggregating person-days, which gave equivalent results to the linear probability model (LPM) method in the previous analysis but significantly reduced computational burden. The second method, m-out-of-n random forests (moonRF), implemented scaling random forests that captured all possible interactions and nonlinearities in the GPS model. To minimize confounding bias, we additionally controlled relative humidity and health care utilizations that were not included previously. Further, we performed low-level analysis by restricting to person-days with exposure levels below increasingly stringent thresholds.

**Results:** We found consistent results between LPM/WLS and moonRF: all exposures were positively associated with mortality rate, even at low levels. For long-term PM<sub>2.5</sub> and O<sub>3</sub>, the effect estimates became larger at lower levels. Long-term exposure to PM<sub>2.5</sub> posed the highest risk: 1 µg/m<sup>3</sup> increase in long-term PM<sub>2.5</sub> was associated with 1053 (95% confidence interval [CI]: 984, 1122; based on LPM/WLS methods) or 1058 (95% CI: 988, 1127; based on moonRF method) early deaths each year among the Medicare population in Massachusetts.

**Conclusions:** This study provides more rigorous causal evidence between PM<sub>2.5</sub>, O<sub>3</sub>, and NO<sub>2</sub> exposures and mortality, even at low levels. The largest effect estimate for long-term PM<sub>2.5</sub> suggests that reducing PM<sub>2.5</sub> could gain the most substantial benefits. The consistency between LPM/WLS and moonRF suggests that there were not many interactions and high-order nonlinearities. In the big data context, the proposed methods will be useful for future scientific work in estimating causality on an additive scale.

**Keywords:** Causality, Air pollution, Generalized propensity score, Data aggregation, Scaling random forests

\* Correspondence: [weiyg@harvard.edu](mailto:weiyg@harvard.edu)

<sup>1</sup>Department of Environmental Health, Harvard T.H. Chan School of Public Health, Landmark Center 4th West, 401 Park Drive, Boston, MA 02215, USA  
Full list of author information is available at the end of the article



© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

## Introduction

Ambient fine particulate matter (PM<sub>2.5</sub>), ozone (O<sub>3</sub>), and nitrogen dioxide (NO<sub>2</sub>) are considered leading causes of death worldwide, largely based on associational studies using traditional statistical methods [1–6]. However, such associations do not necessarily indicate causality [7]. Although a growing body of literature has reported the effect of PM<sub>2.5</sub> exposure on mortality using causal modeling approaches [8–10], few studies so far have examined O<sub>3</sub> and NO<sub>2</sub> [11]. Clearly, O<sub>3</sub> and NO<sub>2</sub> have received less attention than they deserve; long-term O<sub>3</sub> concentration has not been regulated by the U.S. National Ambient Air Quality Standards (NAAQS) and the regulations for NO<sub>2</sub> has been unchanged for decades [12].

Although epidemiologic researchers often report long- or short-term effect of an individual air pollutant, there has been evidence that concurrent air pollution exposures may confound the health effect among each other [2, 3, 13]. In the case of causal analysis, simultaneous assessment of concurrent air pollutants is necessary as it 1) accounts for mutual confounding and thus reduces confounding bias, and 2) allows for comparing the individual effects and identifying the component that is responsible for substantial morbidity and mortality. Indeed, targeting the most harmful air pollutant and its major emission sources based on scientific evidence is the key to efficient and effective air quality regulations [14].

Most air pollution epidemiology studies are conducted using multiplicative models, such as log-linear or Cox proportional hazards models [15]. Such models inherently estimate the effect of exposure on multiplicative scales, which describe the relative change in a health outcome between different exposure levels. In many circumstances, however, it is preferable to measure the absolute effect of exposure on the occurrence of outcome [16]. For example, estimating the additive effect of an air pollutant exposure on mortality rate would give us the number of early deaths due to air pollution, which provides a better sense of the actual size of the health risk and is precisely the type of evidence that U.S. Environmental Protection Agency prefers [17]. In addition, additive models make interaction terms (or their absence) more interpretable, which can help assess effect modification and environmental justice [18].

Recently, we used a parametric generalized propensity score (GPS) adjustment approach to simultaneously estimate causal effects of long- and short-term exposures to PM<sub>2.5</sub>, O<sub>3</sub>, and NO<sub>2</sub> on mortality rate among Medicare beneficiaries in Massachusetts during 2000–2012 [19]. We considered a counting process for analyzing individual survival data [20]. For each exposure, we estimated the GPS at the observed exposure level on each person-day given the other concurrent exposures and all measured confounders. By modeling the binary outcome of

death with linear probability model (LPM), we estimated the additive effect of each exposure on mortality rate. The analysis addressed a critically important question of how many early deaths were caused by air pollution, under the assumption that both GPS model and outcome regression model were correctly specified. However, the GPS models did not capture potential interactions and high-order nonlinearities, making the estimates vulnerable to insufficient confounding control. Further, the counting process data structure with person-day representations of follow-up produces a massive volume of dataset: the whole set of data is comprised of 3.8 billion observations which is about 2 TB in size in the RDS file format, making the analysis computationally expensive.

Using the same cohort, here we proposed two new GPS-based approaches with the goals of increasing computational efficiency and model flexibility in assessing the additive effects of air pollution exposures on mortality. The first approach leveraged the large volume of data by aggregating person-days, which gave the equivalent results to the approach we used in the previous analysis but significantly reduced the computational burden. Building upon the aggregated dataset, the second approach implemented a scaling random forests (RF) method, which increased the flexibility of the GPS model by capturing interactions and nonlinearities. To minimize confounding bias, we also controlled additional community-level confounders that have been suggested as potential confounders. The findings of this analysis will increase the robustness of the association and the validity of causal interpretation of the relationship between air pollution and mortality. In the big data context, the proposed approaches will benefit future scientific work.

## Methods

### Data sources

#### Medicare data

We obtained Medicare enrollment records between January 1, 2000 and December 31, 2012 for beneficiaries aged 65 years and above residing in Massachusetts from the Centers for Medicare and Medicaid Services. We constructed an open cohort with person-day representations of follow-up in which each individual was followed from the maximum of January 1, 2000 or the date of enrollment until death or censoring, whichever occurred earlier. For each beneficiary, we extracted their sex, race/ethnicity, age, Medicaid eligibility, ZIP Code of residence and its latitude and longitude, year of initial enrollment, and date of death if occurred during 2000–2012. Age, Medicaid eligibility, and ZIP Code of residence were updated annually. The outcome of interest is all-cause mortality.

### Exposure assessment

The daily concentrations of ambient PM<sub>2.5</sub>, O<sub>3</sub>, and NO<sub>2</sub> at 1 km × 1 km grid cells across the contiguous US were predicted using geographically weighted regressions that ensembled predictions from RF, gradient boosting, and neural network, which integrated multiple data sources including satellite data, land-use variables, monitoring data, chemical transport model simulations, etc. 10-fold cross-validations on held-out monitors indicated good predictive performance, with mean R<sup>2</sup> of 0.86 for daily PM<sub>2.5</sub>, 0.86 for daily O<sub>3</sub>, and 0.79 for daily NO<sub>2</sub>. Details are published elsewhere [21–23]. The high-resolution and well-validated predictions at 1 km × 1 km grid cells allow us to estimate exposures levels at ZIP Codes with higher degree of accuracy. Using the ZIP Code polygon data generated by Environmental Systems Research Institute [24], for each air pollutant we estimated its daily concentrations in a ZIP Code by averaging the 1 km-gridded predictions with those centroids fall within the boundary of that ZIP Code.

We considered six exposures: long- and short-term exposures to PM<sub>2.5</sub>, O<sub>3</sub>, and NO<sub>2</sub>. For each person on each day, the long-term exposures were defined as annual moving averages of the daily concentrations in the person's ZIP Code of residence (lag 0–364), and the short-term exposures were defined as two-day moving averages of the daily concentrations in the person's ZIP Code of residence (lag 0–1). Following the previous literature, the analysis for short-term O<sub>3</sub> was restricted to person-days in warm season from April to September [3]. The analyses for the other exposures were performed over the entire study period. These exposures were assigned to each person on each day of follow-up.

### Covariates

We made decisions for confounding selection based on both substantive knowledge and the existing literature [2, 25]. Individual-level covariates, including sex (male or female), race (White, Black, or Other), age group (65–69, 70–74, 75–79, 80–84, or ≥ 85 years), and Medicaid eligibility (as a marker of socioeconomic status), were obtained from the Medicare enrollment records. Daily meteorological covariates, including air surface temperature, dew point temperature, and relative humidity with a resolution of 32 km × 32 km, were obtained from the National Centers for Environmental Prediction/National Center for Atmospheric Research datasets, and were matched to each admission based on the latitude and longitude of the centroid of that person's ZIP Code of residence [26]. ZIP Code Tabulation Area (ZCTA)-level socioeconomic and housing characteristics of each year, including median household income, median house value, percent of owner-occupied homes, percent of population living in poverty, percent

of population below high school education, population density, percent of Blacks, and percent of Hispanics, were linearly interpolated between US Census 2000 and 2010 and were extracted from the American Community Survey for years after 2010, and were matched to each admission based on ZCTA to ZIP Code crosswalks [27]. County-level behavioral factors of each year, including percent of ever smokers, lung cancer rate, and average BMI, were obtained from Behavioral Risk Factor Surveillance System, and were linked to each admission based on the ZIP Code of residence [28]. From the Dartmouth Atlas Project [29], we obtained health care utilization variables including percent of persons over age 65 with an annual hemoglobin A1c test, an annual low-density lipoprotein test, and an annual eye exam in each hospital catchment area, and linked them to each admission based on the ZIP Code of residence. The relative humidity and health care utilization variables were not included in our previous analysis. Because they may confound the association, we added them in the current analysis [30, 31].

### GPS methods for causal Modeling

GPS is a powerful tool for confounding control and is increasingly being used in observational studies [32]. In this section, we presented three GPS-based approaches for assessing the additive effects of long- and short-term exposures to PM<sub>2.5</sub>, O<sub>3</sub>, and NO<sub>2</sub> on mortality rate. First we reviewed the LPM approach that was used in the previous analysis. Then we presented two new approaches that were developed upon the LPM: weighted least squares (WLS) and m-out-of-n random forests (moonRF). Each approach consisted of two stages: a design stage where GPS were estimated at the observed exposure levels given all measured confounders, and an analysis stage where the additive effects of exposures on mortality rate were estimated conditional upon the estimated GPS [33].

#### Linear probability model (LPM)

We allowed for time-varying covariates by creating a counting process data structure in which each record represents a person-day of follow-up, indexed by  $i$ . In the design stage, for each exposure we constructed the GPS by fitting a linear regression of the observed exposure level  $T_i$  against column vector of covariates  $C_i$ :

$$T_i = C_i' \beta + \varepsilon_i, \quad (1)$$

where  $\varepsilon_i \sim N(0, \sigma^2)$  under the normality assumption, superscript denoted transpose, and  $\beta$  and  $\sigma$  were estimated by ordinal least squares (OLS). The covariate vector  $C_i$  includes the other five exposures, the individual characteristics (sex, race, 5-year age group, and Medicaid

eligibility), long- (lag 0–364) and short-term (lag 0–1, lag 2–6, and lag 7–12) moving averages of meteorological variables, the community-level socioeconomic and behavioral variables as mentioned earlier, and calendar year for person-day  $i$ . Including the other five exposures controlled for jointly confounded exposures; including long- and short-term meteorological variables controlled for confounding of changing weather and climate; and including calendar year controlled for other confounding by time trends. For short-term exposures, we also included calendar month and day of week to control for seasonal confounding. All the continuous covariates were modeled with cubic polynomials to account for potential nonlinearity. A full list of covariates is provided in Section 1 of Additional file 1.

Given the observed exposure  $T_i$  and covariates  $C_i$ , we estimated the GPS for person-day  $i$  according to Hirano and Imbens [34]:

$$\hat{R}_i = \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \exp\left(-\frac{1}{2\hat{\sigma}^2}(T_i - \hat{T}_i)^2\right), \quad (2)$$

where  $\hat{T}_i = C_i'\hat{\beta}$ . Assuming that the GPS regression (Eq. 1) had been correctly specified,  $\hat{R}_i$  was an estimator of  $R_i$ , which provided a scalar summary of bias introduced by all measured confounders and therefore could be used for confounding control through adjustment.

In the analysis stage, for each exposure we fitted an LPM of binary outcome of death  $Y_i$  against  $T_i$  and  $R_i$ :

$$Y_i = \alpha_0 + \alpha_1 T_i + \alpha_2 R_i + \tau_i, \quad (3)$$

where  $\alpha_0$ ,  $\alpha_1$ , and  $\alpha_2$  were estimated by OLS given  $\hat{R}_i$  estimated from Eq. 2. Assuming that the GPS model (Eq. 1) and the outcome regression model (Eq. 3) had been correctly specified, the OLS estimate  $\hat{\alpha}_1$  was an unbiased causal effect estimate, which can be interpreted as the average difference in mortality rate attributed to each unit increase in the exposure. Such causal interpretation comes from the use of GPS and the collapsibility of LPM, making conditional and marginal estimates numerically the same [35, 36]. Due to the heteroscedastic nature of LPM's residuals [37], we constructed a robust confidence interval (CI) for  $\hat{\alpha}_1$  using sandwich standard error estimates.

#### Weighted least squares (WLS)

One of the main disadvantages of the LPM method is the challenge in processing the massive dataset with person-day representations of follow-up contributed by the Medicare cohort. Here we proposed the WLS method to reduce the computational burden. The WLS aggregated person-days yet retained all the information after the aggregation. As a result, the WLS gave us the

same effect estimates as the LPM but with a significantly improved computational efficiency.

In the design stage, we aggregated the person-days that had the same sex, race, age, Medicaid eligibility, ZIP Code of residence, and date as a single record and assigned the numbers of person-days for that record as weight. This is because the aggregated person-days are identical in terms of all the exposures and covariates, therefore can be treated interchangeably in the analysis. With the aggregated dataset, for each exposure, we fitted a weighted linear regression of the observed exposure level against all the covariates, with continuous ones modeled with cubic polynomials, and estimated the GPS using Eq. 2. We can show that estimating the WLS regression gave the equivalent estimates  $\hat{\beta}$  as estimating Eq. 1 using OLS (Section 2 of Additional file 1).

Person-days with the same exposures, covariates, and thus the estimated GPS may have different outcomes of death (0 or 1). In the analysis stage, we calculated the average outcome for each aggregated person-day group and assigned it to the person-day in the aggregated dataset. For each exposure, with the aggregated dataset, we fitted a weighted linear regression of the averaged outcome against the observed exposure level and the estimated GPS. Similarly, estimating this WLS regression gave the equivalent estimates  $\hat{\alpha}$  as estimating Eq. 3 (Section 2 of Additional file 1). Hence the WLS produced the same effect estimate  $\hat{\alpha}_1$  as the LPM.

In our dataset, because most person-days were identical in terms of the exposures and confounders and therefore were dropped after aggregation, the WLS method saved a lot of storage capacity and significantly speeded up the computation; the number of person-days reduced from 3.8 billion to 60 million after data aggregation and compared with the LPM, the computing time reduced from 3 weeks to 2 days.

#### M-out-of-n random forests (moonRF)

RF is a nonparametric learning method for classification or regression which automatically and thoroughly consider possible nonlinear relationship and interactions. They build individual decision trees through intensive resampling and generally yield better predictive performance than linear regression [38]. In the big data context, Bickel et al. [39] proposed a m-out-of-n bootstrap scheme aiming at addressing the computational burden of standard bootstrapping and proved its consistency. The m-out-of-n bootstrap proceeds by resampling  $m$  observations out of the original dataset (1, ...,  $n$ ) without replacement, where  $m \ll n$ . The number of  $m$  can be as small as  $n^{0.5}$ , much smaller than the typical size of standard bootstrap samples. Setting the number of bootstrap samples at 50 to 100 obtains fairly good predictive



performance, and increasing the number of samples greater than 100 can lead to negligible improvements [38]. Here we estimated the GPS with this idea adopted in the implementation of RF in the design stage of the analysis. The advantage of moonRF over the LPM and WLS is that it has more flexibility to capture any possible interactions and nonlinearities, making the estimates robust to any observed confounding bias [40].

In the design stage, we used the number of person-days aggregated for each record in the aggregated dataset as frequency weight and sampled 62,000 person-days (i.e.,  $N^{0.5}$ , where  $N = 3.8$  billion) without replacement. With this sample, we built a tree for each exposure and made prediction of the exposure for each person-day in the aggregated dataset. We repeated this routine for 100 times. The final predicted exposure level  $\hat{T}_j$  for person-day  $j$  was obtained by averaging the predictions of the 100 trees:

$$\hat{T}_j = \frac{1}{100} \sum_{l=1}^{100} \hat{T}_{jl}. \quad (4)$$

Then applying Eq. 2, we estimated the GPS for each person-day in the aggregated dataset.

In the analysis stage, following the WLS method, we fitted a weighted regression of the averaged outcome against the observed exposure level and the estimated GPS using the aggregated dataset to obtain estimator  $\hat{\alpha}_1$ , which, if both the GPS model and the outcome regression model were correctly specified, was the causal estimate for the additive effect of exposure on mortality rate.

To assess the effects of exposures to low levels of ambient air pollutants, for each method we reconducted the analysis but restricted to person-days with exposure levels below increasingly stringent thresholds, including those well below the levels set in the current NAAQS ( $12 \mu\text{g}\cdot\text{m}^{-3}$  for long-term  $\text{PM}_{2.5}$ ,  $35 \mu\text{g}\cdot\text{m}^{-3}$  for short-term  $\text{PM}_{2.5}$ , 70 parts per billion [ppb] for short-term  $\text{O}_3$ , 53 ppb for long-term  $\text{NO}_2$ , and 100 ppb for short-term  $\text{NO}_2$ ; there is no standard for long-term  $\text{O}_3$ ).

For each exposure, we estimated annual number of early deaths and the 95% CI attributed to each unit increase in the exposure by multiplying the additive effect estimate  $\hat{\alpha}_1$  and annual average number of person-days for the study cohort during 2000–2012.

### Sensitivity analyses

We tested the robustness of the main analysis results by conducting sensitivity analyses with respect to the outcome model flexibility (by modeling GPS with cubic polynomial) and the strategy to adjust for seasonality (by including week-of-year and weekday–weekend dummy variables). We also tested the robustness of the moonRF

method by increasing bootstrap sample size (up to 620,000) and the number trees (up to 500).

The computations of this study were performed on the Research Computing Environment, supported by the Institute for Quantitative Social Science both in the Faculty of Arts and Sciences at Harvard University. We used R software (version 3.5.1) [41], “ranger” package (version 0.12.1) [42], and “biglm” package (version 0.9.1) [43] to perform the analysis.

### Results

Table 1 shows the descriptive statistics of study population. There were a total of 1,503,572 Medicare beneficiaries in the study. Among those, 561,193 (37.3%) deaths occurred. The population consists of more females (57.5%), mostly whites (92.2%), and mostly aged 65–74 years when entering the cohort (69.0%). 17.0% of the population enrolled in Medicaid. Table 2 summarizes the exposure levels across all the beneficiaries’ ZIP Codes of residence during 2000–2012. For each pollutant, the average concentration of long- and short-term exposures were similar while the short-term exposure had greater variation. The exposure levels were mostly below the NAAQS. Descriptive statistics and correlation coefficients among the exposures and covariates are provided in Additional file 1.

Figure 1 shows the results of the three methods at exposure levels below increasingly stringent thresholds. The LPM and WLS methods gave equivalent results and were generally consistent with moonRF. According to

**Table 1** Characteristics of the Medicare population in Massachusetts for the years 2000–2012

	<b>N</b>
Population (%)	1,503,572 (100)
Total person-days	3,874,869,248
Person-days after aggregation	60,708,204
Deaths (%)	561,193 (37.3)
Sex	
Female (%)	864,952 (57.5)
Male (%)	638,620 (42.5)
Race	
White (%)	1,386,883 (92.2)
Black (%)	51,978 (3.5)
Other (%)	64,711 (4.3)
Age at cohort entry	
65–74 (%)	1,037,164 (69.0)
75–84 (%)	335,189 (22.3)
≥ 85 (%)	131,219 (8.7)
Enrollment in Medicaid (%)	255,008 (17.0)

**Table 2** Summary statistics of air pollution exposures across the ZIP Codes of Medicare beneficiaries' residence in Massachusetts during 2000–2012

	Long-term PM <sub>2.5</sub> ( $\mu\text{g}\cdot\text{m}^{-3}$ ) <sup>a</sup>	Short-term PM <sub>2.5</sub> ( $\mu\text{g}\cdot\text{m}^{-3}$ ) <sup>b</sup>	Long-term O <sub>3</sub> (ppb) <sup>a</sup>	Short-term O <sub>3</sub> (ppb) <sup>b,c</sup>	Long-term NO <sub>2</sub> (ppb) <sup>a</sup>	Short-term NO <sub>2</sub> (ppb) <sup>b</sup>
Mean $\pm$ SD	9.0 $\pm$ 1.9	8.9 $\pm$ 5.4	37.5 $\pm$ 3.0	32.6 $\pm$ 10.4	20.4 $\pm$ 8.3	20.5 $\pm$ 11.6
Min	3.3	0.1	25.7	7.9	3.2	0.2
5th percentile	5.8	3.0	32.2	27.6	8.6	5.3
25th percentile	7.6	5.2	35.6	36.7	14.1	11.3
Median	9.0	7.5	37.7	43.2	19.2	18.7
75th percentile	10.1	11.2	39.6	49.6	26.1	27.9
95th percentile	12.1	19.4	42.1	61.6	35.0	41.6
Max	16.4	65.3	47.1	116.0	64.6	119.0

<sup>a</sup> Long-term exposure to air pollution was defined as one-year moving average of the exposure level (lag 0–364). <sup>b</sup> Short-term exposure to air pollution was defined as two-day moving average of the exposure level (lag 0–1). <sup>c</sup> Short-term O<sub>3</sub> was summarized during the warm season from April 1 to September 30

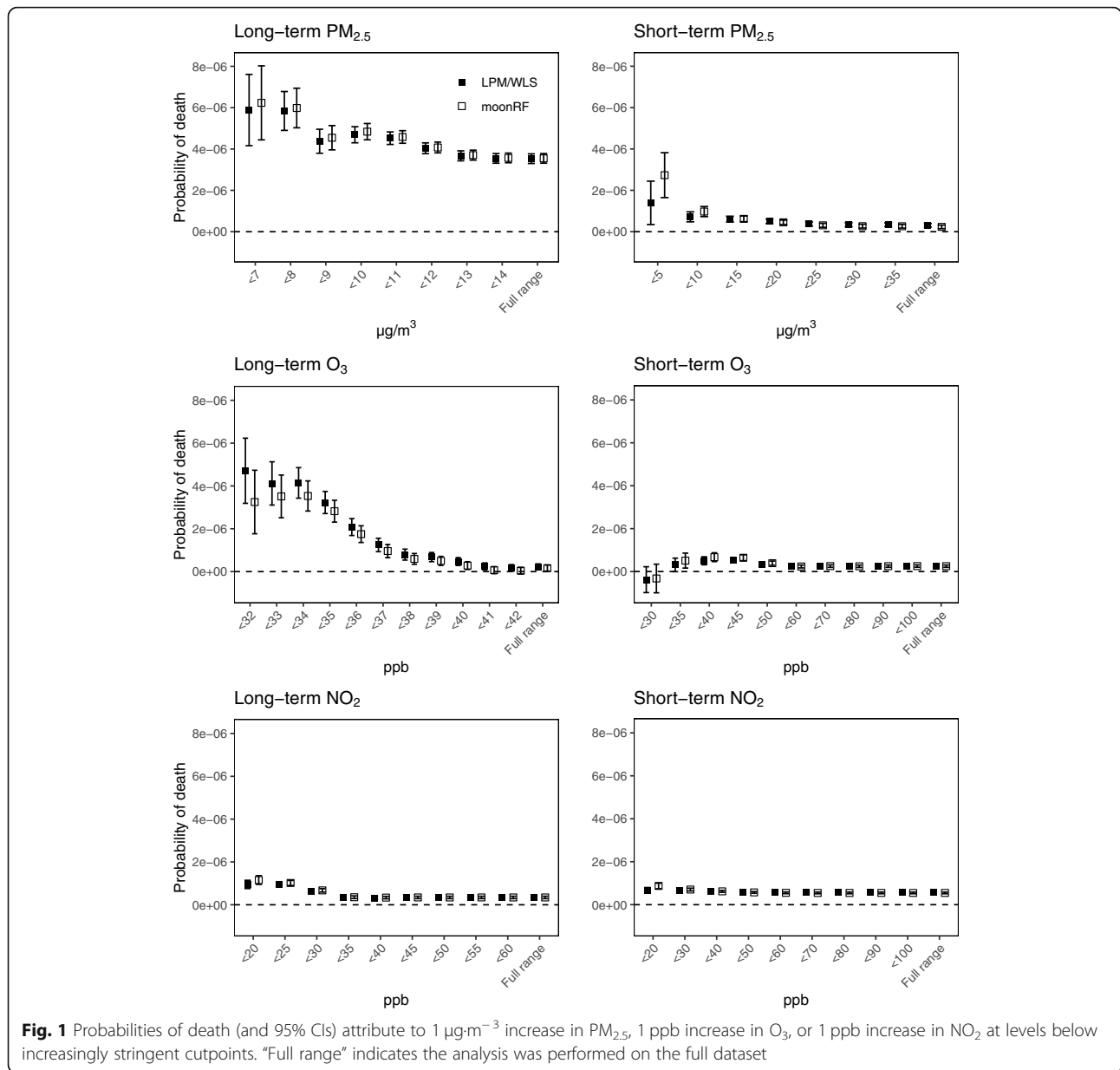
the LPM/WLS, with the full dataset, each 1  $\mu\text{g}\cdot\text{m}^{-3}$  increase in long- and short-term exposures to PM<sub>2.5</sub> was associated with increases of  $3.5 \times 10^{-6}$  (95% CI:  $3.3 \times 10^{-6}$ ,  $3.8 \times 10^{-6}$ ) and  $3.1 \times 10^{-7}$  (95% CI:  $2.2 \times 10^{-7}$ ,  $3.9 \times 10^{-7}$ ) in the probability of death per person-day, respectively; each 1 ppb increase in long- and short-term exposures to O<sub>3</sub> was associated with increases of  $2.2 \times 10^{-7}$  (95% CI:  $0.8 \times 10^{-7}$ ,  $3.6 \times 10^{-7}$ ) and  $2.4 \times 10^{-7}$  (95% CI:  $1.9 \times 10^{-7}$ ,  $3.0 \times 10^{-7}$ ) in the probability of death per person-day, respectively; and each 1 ppb increase in long- and short-term exposures to NO<sub>2</sub> was associated with increases of  $3.3 \times 10^{-7}$  (95% CI:  $2.7 \times 10^{-7}$ ,  $3.8 \times 10^{-7}$ ) and  $5.6 \times 10^{-7}$  (95% CI:  $5.2 \times 10^{-7}$ ,  $6.0 \times 10^{-7}$ ) in the probability of death per person-day, respectively. The moonRF estimates were consistent with those of the LPM/WLS on the full dataset. For long-term PM<sub>2.5</sub> and O<sub>3</sub>, all the methods demonstrated significantly larger effects at lower exposure levels. For short-term O<sub>3</sub>, the LPM/WLS became less consistent with moonRF at very low levels. Numerical values are provided in Section 6 of Additional file 1.

Given the total number of person-days, we estimated the annual number of early deaths due to each exposure (Table 3). With the full dataset, we found that long-term PM<sub>2.5</sub> was associated with the greatest number of early deaths per unit increase in exposure: the annual number of early deaths associated with 1  $\mu\text{g}/\text{m}^3$  increase in long-term exposure to PM<sub>2.5</sub> was 1053 (95% CI: 984, 1122) using LPM/WLS and was 1058 (95% CI: 988, 1127) using moonRF. When restricting the analyses to person-days below the NAAQS, we found greater number of early deaths due to long-term PM<sub>2.5</sub>, short-term PM<sub>2.5</sub>, and short-term O<sub>3</sub>.

The effect estimates remained robust when fitting the outcome regression with GPS modeled by cubic polynomial, including week-of-year and weekday–weekend dummy variables to adjust for seasonality (Section 7 and 8 of Additional file 1), or increasing the bootstrap sample size and the number of trees in moonRF (Section 9 of Additional file 1).

## Discussion

Building upon the LPM method that we used in the previous analysis [19], we proposed two new GPS-based methods, the WLS and moonRF, to estimate the additive effects of long- and short-term exposures to PM<sub>2.5</sub>, O<sub>3</sub>, and NO<sub>2</sub> on mortality rate among Medicare beneficiaries in Massachusetts, 2000–2012, encompassing over 3.8 billion person-days of follow-up. Compared with the LPM, the WLS produced identical results but was superior in computational efficiency, whereas the moonRF was superior in flexibility and bias control. To minimize confounding bias, we additionally adjusted for relative humidity and health care utilization variables, which were not included previously. Our results confirmed previous evidence that all the exposures were significantly associated with mortality rate, even at levels below the current NAAQS. For long-term PM<sub>2.5</sub> and O<sub>3</sub>, the effect sizes were larger when restricting to person-days with exposure levels at increasingly stringent thresholds, suggesting that the exposure-response relationships were nonlinear over full ranges of exposure levels. Using a linear term for each exposure in the outcome regression allowed us to estimate the average difference in mortality rate and further to estimate the number of deaths attributed to a unit increase in the exposure within



**Fig. 1** Probabilities of death (and 95% CIs) attribute to 1  $\mu\text{g}\cdot\text{m}^{-3}$  increase in PM<sub>2.5</sub>, 1 ppb increase in O<sub>3</sub>, or 1 ppb increase in NO<sub>2</sub> at levels below increasingly stringent cutpoints. "Full range" indicates the analysis was performed on the full dataset

different ranges of exposure levels. The additive nature of the estimand provides a clearer measure of the health effects of the exposures and is deemed to be of regulatory interest [17]. Comparing the annual number of early deaths associated with each exposure, we found that the long-term PM<sub>2.5</sub> posed the greatest public health concern, suggesting that reducing PM<sub>2.5</sub> could potentially gain the most substantial health benefits.

The general consistency between the parametric (LPM/WLS) and nonparametric (moonRF) GPS models is a key finding. Such consistency reduces model dependence while increases the internal validity of the use of GPS for summarizing measured confounding [33]. Some studies, including both conventional statistical and

causal modeling analyses, rely on the homogeneity assumption that there are no interaction effects among exposures and confounders [2, 3, 5, 8, 11]. In our study, because the LPM/WLS did not adjust for interactions while the moonRF adjusted for all possible interactions and higher-order nonlinearities, the consistency between the LPM/WLS and moonRF suggests that the homogeneity assumption is likely to hold. In addition, it also suggests that modeling continuous covariates with up to cubic polynomials is sufficient to capture nonlinearities. For long-term O<sub>3</sub>, we found larger difference between the two sets of results at lower levels, which may suggest that the effect was confounded by complex interactions when O<sub>3</sub> formation was inhibited by lower temperature

**Table 3** Annual number of early deaths (and 95% CIs) attribute to 1  $\mu\text{g}\cdot\text{m}^{-3}$  increase in  $\text{PM}_{2.5}$ , 1 ppb increase in  $\text{O}_3$ , or 1 ppb increase in  $\text{NO}_2$ 

		LPM/WLS	moonRF
Full-range analysis <sup>a</sup>	Long-term $\text{PM}_{2.5}$ ( $\mu\text{g}\cdot\text{m}^{-3}$ )	1053 (984, 1122)	1058 (988, 1127)
	Short-term $\text{PM}_{2.5}$ ( $\mu\text{g}\cdot\text{m}^{-3}$ )	92 (67, 117)	69 (44, 95)
	Long-term $\text{O}_3$ (ppb)	66 (24, 107)	48 (6, 90)
	Short-term $\text{O}_3$ (ppb)	73 (57, 89)	74 (58, 91)
	Long-term $\text{NO}_2$ (ppb)	97 (80, 113)	102 (86, 119)
	Short-term $\text{NO}_2$ (ppb)	167 (156, 179)	163 (151, 174)
Below-standard analysis <sup>b</sup>	Long-term $\text{PM}_{2.5}$ ( $\mu\text{g}\cdot\text{m}^{-3}$ )	1203 (1126, 1280)	1214 (1137, 1292)
	Short-term $\text{PM}_{2.5}$ ( $\mu\text{g}\cdot\text{m}^{-3}$ )	101 (74, 127)	78 (52, 105)
	Long-term $\text{O}_3$ (ppb)	NA	NA
	Short-term $\text{O}_3$ (ppb)	100 (74, 127)	116 (87, 145)
	Long-term $\text{NO}_2$ (ppb)	97 (80, 113)	102 (85, 119)
	Short-term $\text{NO}_2$ (ppb)	168 (156, 179)	163 (151, 174)

<sup>a</sup> Full-range analysis was performed on the full dataset. <sup>b</sup> Below-standard analysis was performed on person-days with exposure levels below the NAAQS (< 12  $\mu\text{g}\cdot\text{m}^{-3}$  for long-term  $\text{PM}_{2.5}$ , < 35  $\mu\text{g}\cdot\text{m}^{-3}$  for short-term  $\text{PM}_{2.5}$ , < 70 ppb for short-term  $\text{O}_3$ , < 50 ppb for long-term  $\text{NO}_2$ , and < 100 ppb for short-term  $\text{NO}_2$ ; there is no standard for long-term  $\text{O}_3$ )

[22]. It is also possible that  $\text{NO}_2$  was acted as a surrogate as it was inversely related to  $\text{O}_3$  for long-term exposures, and different methods varied in their ability to identify their effects [44]. However, the lack of supporting evidence requires further studies to address this question.

The additive effect estimates provide evidence of the causal relationship between major air pollutants and mortality, which relied on two key assumptions: no unmeasured confounding and positivity [34]. In the observation setting, these two assumptions must always be made to make appropriate causal inference of any public health problems. For the assumption of no unmeasured confounding, although it is impossible to test whether there exists any unobserved confounding, comparing the results with previous literature provide insights into the validity of this assumption. Using a difference-in-difference approach, Wang et al. estimated that a unit increase in annual  $\text{PM}_{2.5}$  was associated 1.7% increase in mortality rate for people  $\geq 65$  years old in New Jersey [10]. Assuming that the baseline mortality was about 5% for the population, their estimate was equivalent to an additive increase of  $2.3 \times 10^{-6}$  in the probability of death per person-day, which was consistent with our estimates ( $3.5 \times 10^{-6}$ ). Such consistency suggests that our long-term effect estimate of  $\text{PM}_{2.5}$  was not significantly confounded by time-invariant or slowly varying confounders, such as smoking and obesity, since those confounders had been adjusted by design. Similarly, the consistency with a national analysis of short-term  $\text{PM}_{2.5}$  and  $\text{NO}_2$  with the use of negative exposure control provided additional protection against unmeasured confounding [11]. For the positivity assumption, we cannot prove the lack of positivity with the observed data. Consequently, we categorized each exposure by the lower

and upper percentiles and found similar distributions of the estimated GPS across the exposure groups, which suggests that the positivity assumption is likely to hold (Section 10 of Additional file 1). Overall, the consistency with previous studies and the similarity of categorized exposure groups increase the validity of no-unmeasured-confounding and positivity assumptions and, thus, the likelihood of causal connections between the major air pollutants and mortality.

The proposed GPS adjustment approaches have several advantages. First, the use of GPS allows us to adjust for a large number of confounders of both long- and short-term exposures and adequately control for potential nonlinearities and interactions. Because the objective of propensity score estimation is to obtain the best predictive accuracy, we do not need to concern about over-parameterizing. Second, in the analysis stage, the small outcome model with only two covariates, the exposure and the estimated GPS, makes the model fitting and generating robust CIs substantially efficient. Third, the use of OLS regression in the analysis stage also provides a causal interpretation of the exposure coefficient, which comes from the fact that the conditional and marginal estimates are numerically the same.

This study also has limitations. First, although air pollution levels were estimated from models with excellent out-of-sample prediction ability, there is likely measurement error when exposure levels were averaged and assigned to ZIP Codes, which may attenuate effect estimates [45]. While upward bias is also possible, it relies on a combination of large exposure error and high exposure correlation with omitted confounders, which we believe is unlikely [46]. Second, we were not able to adjust for the history of chronic diseases because such



information is not available for the Medicare enrollment records, which may leave residual confounding. Third, information bias inherent to the lack of individual-level data, apart from age, sex, race, and Medicaid eligibility, could be present given that ZIP Code was the finest geographical unit we could use to link covariates with each beneficiary.

## Conclusions

Considering the internal validity of the design process for the estimation of GPS and the consistency with previous literature that use several different strategies to address confounding, this study provides more rigorous evidence of the causal relationships between long- and short-term exposures to PM<sub>2.5</sub>, O<sub>3</sub>, and NO<sub>2</sub> and mortality, even at levels below the current NAAQS. The general consistency between the parametric LPM/WLS and nonparametric moonRF methods suggests that there were not many interaction effects among confounders, and that modeling continuous covariates with up to cubic polynomials was sufficient to capture nonlinearities. In the big data context, the proposed GPS-based methods will be useful in estimating causality on an additive scale for the future scientific work.

## Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12940-021-00704-3>.

### Additional file 1.

## Abbreviations

PM<sub>2.5</sub>: Ambient fine particulate matter; O<sub>3</sub>: Ozone; NO<sub>2</sub>: Nitrogen dioxide; GPS: Generalized propensity score; NAAQS: National Ambient Air Quality Standards; LPM: Linear probability model; WLS: Weighted least squares; moonRF: m-out-of-n random forests; ppb: Parts per billion

## Acknowledgements

Not applicable.

## Authors' contributions

Y.W. designed research, performed analysis, and wrote the paper; B.C., P.K., J.Y., L.L., A.Z., and J.S. prepared data and designed research. All authors helped interpret the results and provided comments.

## Funding

This study was supported by National Institutes of Health (NIH)/ National Institute of Environmental Health Sciences (NIEHS) grants P30 ES000002, R01 ES024332; the United States Environmental Protection Agency (US EPA) grants RD-83587201, RD-83615601, and RD-83587201. Its contents are solely the responsibility of the grantee and do not necessarily represent the official views of the US EPA. Furthermore, the US EPA does not endorse the purchase of any commercial products or services mentioned in the publication.

## Availability of data and materials

The exposure data during the current study are available from the corresponding author on reasonable request. The Medicare data are available upon request to the Centers for Medicare and Medicaid Services. The other data are publicly available, with sources described in the manuscript.

## Ethics approval and consent to participate

This study was approved by the institutional review board at the Harvard T.H. Chan School of Public Health and was exempt from informed consent requirements as a study of previously collected administrative data.

## Consent for publication

Not applicable.

## Competing interests

Dr. Joel Schwartz serves as an expert witness for the United States Department of Justice in a case involving a Clean Air Act violation.

## Author details

<sup>1</sup>Department of Environmental Health, Harvard T.H. Chan School of Public Health, Landmark Center 4th West, 401 Park Drive, Boston, MA 02215, USA. <sup>2</sup>Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA. <sup>3</sup>Department of Biostatistics, School of Public Health, Brown University, Providence, RI, USA. <sup>4</sup>Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA.

Received: 27 October 2020 Accepted: 16 February 2021

Published online: 23 February 2021

## References

- Dockery DW, Pope CA, Xu X, Spengler JD, Ware JH, Fay ME, et al. An association between air pollution and mortality in six U.S. cities. *N Engl J Med*. 1993;329(24):1753–9.
- Di Q, Dominici F, Schwartz JD. Air pollution and mortality in the Medicare population. *N Engl J Med*. 2017;377(15):1498–9.
- Di Q, Dai L, Wang Y, Zanobetti A, Choirat C, Schwartz JD, et al. Association of Short-term Exposure to air pollution with mortality in older adults. *Jama*. 2017;318(24):2446–56.
- Wei Y, Wang Y, Di Q, Choirat C, Wang Y, Koutrakis P, et al. Short term exposure to fine particulate matter and hospital admission risks and costs in the Medicare population: time stratified, case crossover study. *BMJ*. 2019; 367:l6258.
- Jerrett M, Burnett RT, Pope CA 3rd, Ito K, Thurston G, Krewski D, et al. Long-term ozone exposure and mortality. *N Engl J Med*. 2009;360(11):1085–95.
- Faustini A, Rapp R, Forastiere F. Nitrogen dioxide and mortality: review and meta-analysis of long-term studies. *Eur Respir J*. 2014;44(3):744–53.
- Lilienfeld DE. Definitions of epidemiology. *Am J Epidemiol*. 1978;107(2):87–90.
- Wang Y, Lee M, Liu P, Shi L, Yu Z, Abu Awad Y, et al. Doubly Robust Additive Hazards Models to Estimate Effects of a Continuous Exposure on Survival. *Epidemiology (Cambridge, Mass)*. 2017;28(6):771–9.
- Schwartz J, Austin E, Bind MA, Zanobetti A, Koutrakis P. Estimating causal associations of fine particles with daily deaths in Boston. *Am J Epidemiol*. 2015;182(7):644–50.
- Wang Y, Kloog I, Coull BA, Kosheleva A, Zanobetti A, Schwartz JD. Estimating causal effects of Long-term PM<sub>2.5</sub> exposure on mortality in New Jersey. *Environ Health Perspect*. 2016;124(8):1182–8.
- Schwartz J, Fong K, Zanobetti A. A National Multicity Analysis of the causal effect of local pollution, [formula: see text], and [formula: see text] on mortality. *Environ Health Perspect*. 2018;126(8):87004.
- Agency USEP. National Ambient Air Quality Standards Table 2016 [Available from: <https://www.epa.gov/criteria-air-pollutants/naaqs-table>].
- Dominici F, Peng RD, Barr CD, Bell ML. Protecting human health from air pollution: shifting from a single-pollutant to a multipollutant approach. *Epidemiology (Cambridge, Mass)*. 2010;21(2):187–94.
- Thakrar SK, Balasubramanian S, Adams PJ, IsML A, Muller NZ, Pandis SN, et al. Reducing mortality from air pollution in the United States by targeting specific emission sources. *Environmental Science and Technology Letters*. 2020;7(9):639–45.
- Krzyzanowski M, Cohen A, Anderson R, Group WHOW. Quantification of health effects of exposure to air pollution. *Occup Environ Med*. 2002;59(12): 791–3.
- Noordzij M, van Diepen M, Caskey FC, Jager KJ. Relative risk versus absolute risk: one cannot be interpreted without the other. *Nephrol Dial Transplant*. 2017;32(suppl\_2):iii13–i8.
- Owens EO, Patel MM, Kirrane E, Long TC, Brown J, Cote I, et al. Framework for assessing causality of air pollution-related health effects for reviews of

- the National Ambient air Quality Standards. *Regul Toxicol Pharmacol.* 2017; 88:332–7.
18. De Gonzalez A, Cox D. Interpretation of interaction: A review. 2007;1(2):371–85.
  19. Wei Y, Wang Y, Wu X, Di Q, Shi L, Koutrakis P, et al. Causal effects of air pollution on mortality in Massachusetts. *Am J Epidemiol.* 2020.
  20. Fleming TR, Harrington DP. *Counting Processes and Survival Analysis.* 2nd ed: Wiley; 2013.
  21. Di Q, Amini H, Shi L, Kloog I, Silvern R, Kelly J, et al. An ensemble-based model of PM<sub>2.5</sub> concentration across the contiguous United States with high spatiotemporal resolution. *Environ Int.* 2019;130:104909.
  22. Requia WJ, Di Q, Silvern R, Kelly JT, Koutrakis P, Mickley LJ, et al. An Ensemble Learning Approach for Estimating High Spatiotemporal Resolution of Ground-Level Ozone in the Contiguous United States. *Environmental science & technology.* 2020.
  23. Di Q, Amini H, Shi L, Kloog I, Silvern RF, Kelly JT, et al. Assessing NO<sub>2</sub> concentration and model uncertainty with high spatiotemporal resolution across the contiguous United States using ensemble model averaging. *Environmental science & technology.* 2019;54(3):1372–84.
  24. Institute ESR. *Esri Data & Maps 10. Redlands: An Esri White Paper;* 2010.
  25. VanderWeele TJ. Principles of confounder selection. *Eur J Epidemiol.* 2019; 34(3):211–9.
  26. Kalnay E, Kanamitsu M, Kistler R, Collins W, Deaven D, Gandin L, et al. The NCEP/NCAR 40-year reanalysis project. *Bull Am Meteorol Soc.* 1996;77(3): 437–72.
  27. Mapper U. ZIP code to ZCTA crosswalk; 2014.
  28. CDC. Behavioral Risk Factor Surveillance System Survey Questionnaire. Atlanta, Georgia: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention; 2004.
  29. Cronenwett JL, Birkmeyer JD. The Dartmouth atlas of vascular health care. *Cardiovasc Surg.* 2000;8(6):409–10.
  30. Barreca AI. Climate change, humidity, and mortality in the United States. *J Environ Econ Manage.* 2012;63(1):19–34.
  31. Lee KC, Sturgeon D, Lipsitz S, Weissman JS, Mitchell S, Cooper Z. Mortality and health care utilization among Medicare patients undergoing emergency general surgery vs those with acute medical conditions. *JAMA Surg.* 2020;155(3):216–23.
  32. Zhao S, van Dyk DA, Imai K. Propensity score-based methods for causal inference in observational studies with non-binary treatments. *Stat Methods Med Res.* 2020;29(3):709–27.
  33. Rubin DB. For objective causal inference, design trumps analysis. *Ann Appl Stat.* 2008;2(3):808–40.
  34. Hirano K, Imbens GW. The Propensity Score with Continuous Treatments. In: Gelman A, Meng XL, editors. *Applied Bayesian Modeling and Causal Inference from Incomplete-Data.* Hoboken, NJ: John Wiley & Sons, Ltd; 2004. p. 73–84.
  35. Vellaisamy P, Vijay V. Collapsibility of regression coefficients and its extensions. *Journal of Statistical Planning and Inference.* 2008;138(4):982–94.
  36. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology (Cambridge, Mass).* 2000;11(5): 550–60.
  37. Battey HS, Cox DR, Jackson MV. On the linear in probability model for binary data. *R Soc Open Sci.* 2019;6(5):190067.
  38. Genuera R, Poggib J-M, Malot C, Vialaneix NV. Random Forests for Big Data. *Big Data Research.* 2017;9:28–46.
  39. Bickel PJ, Bickel F, Zwet WRV. Resampling fewer than n. observations: gains, losses, and remedies for losses. *Stat Sin.* 1997;7(1):1–31.
  40. Dudik M, Langford J, Li L. Doubly robust policy evaluation and learning. *The international conference on machine learning.* International Machine Learning Society: Bellevue, Washington; 2011.
  41. Team RCD. R: A language and environment for statistical computing. 3.5.1 ed: R Foundation for Statistical Computing; 2010.
  42. Wright M, Wager S, Probst P. Package ‘ranger’: A Fast Implementation of Random Forests. 0.12.1 ed: CRAN; 2020.
  43. Lumley T. Package ‘biglm’: bounded memory linear and generalized linear models. 0.9.1 ed: CRAN; 2020.
  44. Semple DR, Song F, Gao Y. Seasonal characteristics of ambient nitrogen oxides and ground-level ozone in metropolitan northeastern New Jersey. *Atmospheric Pollution Research.* 2012;3(2):247–57.
  45. Weiskopf MG, Webster TF. Trade-offs of Personal Versus More Proxy Exposure Measures in Environmental Epidemiology. *Epidemiology (Cambridge, Mass).* 2017;28(5):635–43.
  46. Butland BK, Samoli E, Atkinson RW, Barratt B, Katsouyanni K. Measurement error in a multi-level analysis of air pollution and health: a simulation study. *Environmental health : a global access science source.* 2019;18(1):13.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

