Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Haomin Li Date

Health Effects of Air Pollutant Mixtures on Overall Mortality Among the Elderly Population Using Bayesian Kernel Machine Regression (BKMR)

By

Haomin Li

Master of Science in Public Health

Epidemiology

_____

Kyle Steenland, Ph.D.
Committee Chair

_____

Liuhua Shi, Sc.D.
Committee Member

Health Effects of Air Pollutant Mixtures on Overall Mortality Among the Elderly Population Using Bayesian Kernel Machine Regression (BKMR)

By

Haomin Li

B.S.
Peking University
2018

Thesis Committee Chair: Kyle Steenland, Ph.D.

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Epidemiology
2021

Abstract

Health Effects of Air Pollutant Mixtures on Overall Mortality Among the Elderly Population Using Bayesian Kernel Machine Regression (BKMR)
By Haomin Li

Background: It is well documented that fine particles matter ($PM_{2.5}$), ozone (O_3), and nitrogen dioxide (NO_2) are associated with a range of adverse health outcomes. However, most epidemiological studies have focused on understanding their additive effects, despite that individuals are exposed to multiple air pollutants simultaneously that are likely correlated with each other.

Method: We applied a novel method - Bayesian Kernel machine regression (BKMR) and conducted a population-based cohort study to assess the individual and joint effect of air pollutant mixtures ($PM_{2.5}$, O_3, and NO_2) on all-cause mortality among the 1,406,185 Medicare population in 15 cities with 656 different ZIP codes in the southeastern US.

Results: The results suggest a strong association between pollutant mixture and all-cause mortality, mainly driven by $PM_{2.5}$. The positive association of $PM_{2.5}$ with mortality appears stronger at lower percentiles of other pollutants. An interquartile range change in PM2.5 concentration was associated with a significant increase in mortality of 1.7 (95% CI: 0.5, 2.9), 1.6 (95% CI: 0.4, 2.7) and 1.4 (95% CI: 0.1, 2.6) standard deviations (SD) when O_3 and NO_2 were set at the 25^{th}, 50^{th}, and 75^{th} percentiles, respectively.

Conclusion: BKMR analysis did not identify statistically significant interactions among $PM_{2.5}$, O_3, and NO_2. However, since the small sub-population might weaken the study power, additional studies (in larger sample size and other regions in the US) are in need to reinforce the current finding.

Health Effects of Air Pollutant Mixtures on Overall Mortality Among the Elderly Population Using Bayesian Kernel Machine Regression (BKMR)

By

Haomin Li

B.S.
Peking University
2018

Thesis Committee Chair: Kyle Steenland, Ph.D.

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Epidemiology
2021

1. Introduction

Ambient air pollution is a major public health concern and estimates to be responsible for 7.6% of all deaths worldwide in 2016 (WHO, 2016). Air pollution is a complex mixture composed of both solid particles (e.g., fine particles matter - $PM_{2.5}$) and gaseous pollutants (e.g., ozone - O_3, nitrogen dioxide - NO_2). $PM_{2.5}$, O_3, and NO_2 account for by far the greatest health burden globally (Fann et al., 2012; Plass et al., 2019; Li et al., 2021). In particular, long-term exposure to these air pollutants is associated with increased morbidity and mortality of various diseases (Faustini et al., 2014; Cohen et al., 2017; Burnett et al., 2018).

Over the past decades, extensive research has focused on the adverse health effects of $PM_{2.5}$ and O_3, and most recently NO_2, with the effect of each pollutant often modelled in isolation (Crouse et al., 2015; Orioli et al., 2018; Barzeghar et al., 2020; Wei et al., 2020). However, in realistic scenarios, the study populations are invariably exposed to a mixture of multiple air pollutants simultaneously. Furthermore, the joint concentration-response (C-R) relationship between the multi-pollutant exposure and the health outcome can be highly nonlinear, and often exhibiting non-trivial patterns of between-pollutant interaction (Bobb et al., 2015). Indeed, recent investigations found evidence of nonlinear associations between long-term exposure to air pollution and mortality and morbidity (Li et al., 2018; Vodonos et al., 2018; Shi et al., 2020; Yu et al., 2020).

To this end, the traditional approaches to pollution mixture modeling have relied on multivariable parametric regression, which estimates the independent effect of each pollutant as linear terms, adjusting for the confounding effect of the other co-pollutants (Brook et al., 2007; Tolbert et al., 2007; Chen et al., 2010). However, if multiple pollutants exhibit a non-trivial correlation structure and complex nonlinear relationship with the

outcome, this approach will be suffer the issue of multicollinearity and model misspecification (Allen, 1997; Wang et al., 2014). On the other hand, nonparametric or latent-variable approaches have also been proposed to better account for the within-mixture correlation structure or the nonlinear effect of the pollutants. The examples include recursive partitioning (Loh, 2011), supervised principal component analysis (Roberts and Martin, 2006), and the latent class analysis (Proust-Lima et al., 2007). However, these methods either rely on explicit assumptions on the functional form of the C-R relationship, or lose statistical efficiency quickly as the data dimension increases (Hastie et al., 2009). Therefore, they are not suitable for modeling a nonlinear, high-dimensional C-R relationship whose function form is *a priori* unknown. Hence, a flexible and yet statistically efficient method is needed to properly account for the nonlinear and interactive health effect among multiple concurrent air pollution exposures, for the purpose of generating rigorous and informative statistical evidence to facilitate evidence-based regulatory decisions.

In this work, we apply the Bayesian kernel machine regression (BKMR), a novel semi-parametric modeling approach to flexibly capture the joint effect of the mixture components, allowing for potential interactions and nonlinear effects. As a statistical model, BKMR offers two appealing advantages when compared to the previous purely parametric or nonparametric approaches. First, it handles the joint effect of multiple pollutants using a kernel machine regression model, thereby capturing the potentially complex and nonlinear joint dose-response curve of multiple exposures while maintaining good statistical power. Second, it allows for the disentangling of the joint effect of pollutant mixture into its main-effect and interactive-effect components while properly accounting for model uncertainty (Bobb et al., 2015). As a result, for the purpose of studying the health effect of pollutant

mixture, BKMR allows researchers to investigate (1) the joint effect of the pollutant mixture as a whole, (2) the individual C-R relationship of each mixture component adjusting for the other pollutants, and finally (3) the relative contribution of each individual pollutant to the overall effect. This can provide a comprehensive view of the statistical structure underlying the effects of multiple pollutants. Recently, there is emerging evidence for employing BKMR to estimate the mixed chemicals and health outcomes, such as systemic autoimmune rheumatic makers, cardiovascular endpoints, and neurodevelopmental outcomes, birth outcomes (Domingo-Relloso et al., 2019; Ashrap et al., 2020; Yin et al., 2020; Zhao et al., 2020). However, the BKMR approach, to the best of our knowledge, has not yet been utilized in studies on air pollution and all-cause mortality.

The health effects of air pollutant mixtures among the elderly population are of particular interest in the southeastern United States (SEUS), because a large fraction of the US elderly that are considered to be most vulnerable to air pollution have moved to this area with mild weather. However, due to the computational limitation of BKMR when dealing with large datasets(Bobb et al., 2015), we focused our study on elderly population in North Carolina, South Carolina, and Georgia, which could represent the older population in the SEUS. In this analysis, we conducted a population-based cohort study of the Medicare beneficiaries (aged 65 or over) in 15 cities in SEUS and applied the BKMR method to estimate the joint effect of three predominant air pollutants (i.e., $PM_{2.5}$, O_3, and NO_2) on all-cause mortality, and to disentangle the health effect of individual pollutants from that of the overall mixture.

2. Methods

2.1. Study Population

We obtained the health data from the Centers for Medicare and Medicaid Services., which contains the information of more than 96% of population aged 65 years or older in the United States. Our study population included 1,406,185 Medicare enrollees residing in 15 cities with 656 different ZIP codes in North Carolina, South Carolina, and Georgia from January 1st, 2000 to December 31st, 2010. The enrollees entered the Medicare cohort when they turned 65, and were followed-up until death, censoring, or the end of the study period. The study outcome was all-cause mortality. For each beneficiary, we extracted individual information on age at baseline, sex, race, Medicaid eligibility (a proxy for low socioeconomic status - SES), ZIP code of residence, and date of death (up to December 31st, 2010) from the Medicare enrollment file. ZIP code of residence and calendar year were further used to link the health records with air pollutant concentrations and covariates. This study was approved by the Institutional Review Board of Emory University and a waiver of informed consent was granted.

2.2. Air Pollution Exposures

We applied previously estimated daily ambient $PM_{2.5}$, O_3, and NO_2 levels from 2000-2016 at 1 km spatial resolution in the contiguous U.S. using well-validated ensemble machine learning model, which integrated multiple predictor variables and three machine learning algorithms (Di et al., 2019b, a; Requia et al., 2020). Briefly, at the first stage, we respectively fit a neural network, random forest, and gradient boosting with more than 100 predictor variables. Predictor variables included satellite-based measurements, meteorological variables, chemical transport model simulations, and land-use variables, and the model was trained at all monitoring stations in the U.S.. Then, for each air pollutant of interest we

combined the pollutant estimates from three machine learners in a generalized additive geographically weighted model and generated the final predictions. This ensemble learning approach was found to achieve excellent model performance, with 10-fold cross-validation R^2 of 0.86, 0.90, and 0.79 for $PM_{2.5}$, O_3, and NO_2, respectively. Based on the daily predictions at 1 km^2 grid cells, we estimated the daily concentrations in a ZIP code by averaging these gridded predictions whose centroids fall within the boundary of a given ZIP code. We further calculated the annual averages for $PM_{2.5}$ and NO_2 as well as the warm-season (May 1 to October 31) average for O_3 from 2000 to 2010 in each ZIP code. For each Medicare enrollee, we assigned the mean of the annual or warm-season average pollutant concentrations across the years they were in the cohort, according to the ZIP code of residence.

2.3. Covariates

We derived daily air temperature and relative humidity data, at a spatial resolution of 32 km × 32 km, across the US from the North American Regional Reanalysis (NARR) from 2000-2010. We matched each ZIP code centroid to the nearest 32 km grid cell, and assigned the daily meteorological data, and then calculated the annual average. Seven ZIP code-level variables, including population density, percent Black, percent of the population with less than a high school degree, percent below the poverty level, median house value, median household income, and percent of owner-occupied housing units were derived from the 2000 U.S. Census and the 2010 U.S. Census. We also obtained county-level variables, smoking rate and body mass index (BMI), from the 2000-2010 Behavioral Risk Factor Surveillance System (BRFSS). These county-level variables were matched to ZIP codes whose centroids fell within the county boundary.

2.4. Statistical Analysis

We modeled the mortality outcome as z-scored. The statistical analysis consisted of two stages: In the first stage, we applied the BKMR method to estimate the city-specific joint-effect of exposure mixtures on mortality; in the second stage, we estimated the global health effect across all cities by pooling the city-specific effect estimates using weighted average ensemble. Medicare enrollees might have more than one observation since the data were recorded per individual per year. To fit the BKMR model, we collapsed the data among each enrollee by extracting the ultimate death information, taking the minimum of the age (i.e., age at enrollment), and calculating the mean of annual or warm-season air pollution exposure levels and the rest covariates over the study period.

2.4.1. Stage I: Estimating City-specific Nonlinear Health Effect using BKMR

To properly capture the potential interaction and nonlinear effects among mixture components, we employed the BKMR method to flexibly model the association between multi-pollutant mixtures and all-cause mortality for each of the 15 cities (Bobb et al., 2015). Specifically, for each individual $i = 1, \ldots, n$, BKMR models the relation between the health outcome Y_i, the background covariates $\mathbf{x}_i$ and the exposures of interest $\mathbf{z}_i$ as:

$$logit(P(Y_i|\mathbf{x}_i, \mathbf{z}_i)) = h(\mathbf{z}_i) + \mathbf{x}_i^\top \boldsymbol{\beta} \quad (1)$$

where Y_i is the health outcome of all-cause mortality, $\mathbf{x}_i$ is a set of p potential confounders (e.g., sex, age, and race) and $\mathbf{z}_i$ is a vector of q pollutant components (i.e., $PM_{2.5}$, O_3, and NO_2). Here $h(\cdot)$ is a flexible nonparametric function that represents the nonlinear C-R exposure-

response relationship that accommodates the interactions among exposures and is using the kernel-machine regression (Liu et al., 2007).

To this end, BKMR estimates the model in Equation (1) via Bayesian inference to account for uncertainty due to estimation of a high-dimensional set of exposures and multiple-testing penalty (Bobb et al., 2015). Briefly, BMKR models the nonlinear function $h(\cdot)$ using a Gaussian process model with a radial basis function (RBF) kernel, and also estimates the individual contribution of each pollutants by placing a spike-and-slab priors onto the pollutant components $\mathbf{z}_i$. The posterior estimation is conducted via Markov chain Monte Carlo (MCMC) sampling. To estimate the contribution of individual pollutant to the health effect of the overall mixture, we followed the recommendation of the BMKR authors to consider two approaches: (1) The marginal nonlinear C-R curve between the individual pollutant and the health outcome, by fixing the health effect of other mixture components at the 25^{th}, 50^{th}, and 75^{th}quantiles, and (2) the *posterior inclusion probability* (PIP), which is the probability that a particular pollutant within the mixture was included in the model by the spike-and-slab variable selector in the posterior sample. More details of kernel-machine regression and PIP are provided in the supplemental materials.

2.4.2. Stage II: Estimating Global Health Effect via Weighted Average Ensemble

Since it is computationally intensive to fit a BKMR model to the entire population in the dataset, we applied an ensemble approach to estimate the global effect by pooling the city-specific health-effect estimates used a weighted average ensemble method. Specifically, we aggregate all the city-specific effect estimates with weights proportional to their number of observations.

2.5. Sensitivity Analysis

We conducted three sensitivity analyses to assess the robustness of our results. First, we omitted a different set of covariates in each model and compared the effect estimates, in order to assess the importance of omitted covariates. Second, we further adjusted for the Normalized Difference Vegetation Index (NDVI), an indicator for surrounding greenness. We obtained the monthly NDVI values at 0.05° (~5 km) resolution from the MODIS satellite and calculated the annual averages based on all covered grids for each ZIP code. Third, we conducted a subgroup analysis by geographical location.

All statistical analyses were conducted using R software, version 3.6.1 and mainly completed by the 'bkmr' package (Bobb et al., 2018), version 0.2.0.

3. Results

3.1. Study Population Characteristics

Our cohort consisted of 1,406,185 Medicare enrollees aged 65 years and older residing in 656 different ZIP codes in the southeastern US. From 2000 to 2010, a total of 416,340 deaths were recorded in this study. The participants were 41.8% male, 75.9% white, and had a mean age of 67.0 (standard deviation, SD=1.4) years at baseline. During the study period, the overall annual mean $PM_{2.5}$, NO_2, and warm-season O_3 concentrations across 15 cities were 13.1 ppb (SD=2.0), 21.6 ppb (SD=7.3), and 47.3 ppb (SD=3.4), respectively. We found significant Pearson correlations (t-test, $P<0.05$) among three pollutants, with positive pairwise correlation coefficients. Specifically, the correlation coefficients are 0.7 between $PM_{2.5}$ and NO_2, 0.8 between $PM_{2.5}$ and O_3, and 0.6 between NO_2 and O_3 (Supplementary Figure

1). Table 1 summarizes the demographic characteristics and average pollutant concentrations in this study.

3.2. BKMR Analysis

We first examine the nonlinear health effect estimate of the overall pollutant mixtures on all-cause mortality. Specifically, in Figure 1, we estimate the posterior mean and associated 95% credible intervals of the estimated change in (z-scored) all-cause mortality when three pollutants were set at a particular percentile compared to when three pollutants were all at their 50^{th} percentile. As shown, we found that the estimated risk of all-cause mortality increased with a simultaneous increase of three pollutants, from 25^{th} percentile to 75^{th} percentile (i.e., an interquartile range [IQR]), as compared to when all pollutants are at their median values (i.e., 50^{th} percentile), indicating a positive joint effect of pollutant mixtures. Particularly, when all three pollutants at or above their 65^{th} percentile, the joint effect of $PM_{2.5}$, O_3, and NO_2 on mortality was statistically significantly different (i.e., its 95% credible intervals do not overlap with zero) than when all three pollutants at their median values.

To disentangle which pollutant dominates the overall effect of the mixtures, we calculated the PIP of the mixture components $PM_{2.5}$, O_3, and NO_2. We observe the pollutant mixture as a whole (i.e., $PM_{2.5}$, O_3, and NO_2) to be strongly associated with all-cause mortality, with the PIP values higher than or equal to 0.5 most of the time. To test the effect of pollutant mixtures rigorously, we performed a hypothesis test (Liu et al., 2007; Deng et al., 2018), with the null hypothesis as no effect of pollutant mixtures. Given the large sample, we subsampled 150 independent datasets, each with 500 individuals without replacement. We performed hypothesis testing in each dataset and obtained 150 p-values. We then combined the results

from these 150 independent tests using Fisher's method and the resulting p-value was less than 10^{-6}. Among all models, 99.3%, 86.7%, and 84.0% of PIPs were higher than the 0.5 thresholds for $PM_{2.5}$, O_3, and NO_2, respectively (Figure 2a). We then sought to figure out the dominant pollutant by changing the threshold of PIP value which we used to determine the variable to be included or not. That is, we considered a pollutant to be important only when its associated PIP is greater than a threshold τ. As shown in Figure 2b, as we increase the threshold τ from 0 to 1, we observed that $PM_{2.5}$ always had the greatest proportion of PIP values that were larger than the threshold, indicating that $PM_{2.5}$ has a stronger explanatory power for all-cause mortality compared with O_3 and NO_2.

We next investigated the importance of the pollutant mixture in contributing to the health outcome by estimating the change in the risk of all-cause mortality associated with an IQR increase in a single pollutant level, while the other pollutants are fixed at 25th, 50th, or 75th percentile levels, respectively. We found that $PM_{2.5}$ is the only pollutant displaying a positive and significant effect in this study (Figure 3). The association between $PM_{2.5}$ and mortality appears stronger at lower percentiles of other pollutants. Specifically, An IQR change in $PM_{2.5}$ concentration is associated with a significant increase in mortality of 1.7 (95%CI: 0.5, 2.9), 1.6 (95%CI: 0.4, 2.7) and 1.4 (95%CI: 0.1, 2.6) SDs when O_3 and NO_2 are set at the 25th, 50th, and 75th percentiles, respectively. In addition, the effect estimates in Figure 3 suggested possible interaction of the pollutant mixture, despite the lack of statistical significance (due to highly overlapping confidence intervals). Specifically, we found that the effects of $PM_{2.5}$ on mortality decreased as NO_2 and O_3 both increased from their 25th to their 75th percentiles.

To further investigate the potential nonlinear C-R relationship and possible interaction of the mixture, we estimated both univariate and bivariate C-R functions. Figure 4

demonstrated the univariate C-R functions and 95% credible intervals (shaded area) for each pollutant with the other pollutants fixed at the median values. We observed a significantly increasing C-R relationship with very tight credible bands for $PM_{2.5}$ within the range 9μg/m^3 to 17 μg/m^3, and the curve flattened at the lower and higher levels with large uncertainties. On the other hand, the data do not support a significant association of O_3 or NO_2 with mortality, due to the large uncertainty in the C-R curves for both O_3 and NO_2. The high credible intervals for NO_2 and O_3 possibly due to weaker exposure estimation since these pollutants are more spatially heterogeneous compared to $PM_{2.5}$.

Finally, we assessed the bivariate C-R functions for the three pollutants to investigate the possible interactions (Figure S2). The slopes for each pollutant are similar at varying levels of the other pollutants, suggesting a lack of statistically significant interaction between individual pollutants. Notice that the gaps between the C-R curves for O_3 or NO_2 were large for different levels of $PM_{2.5}$ (row 1, columns 2 and 3), while the gaps between the C-R curves for $PM_{2.5}$ were small for different levels of O_3 or NO_2. This suggests that $PM_{2.5}$ has stronger association with mortality compared to O_3 or NO_2.

4. Discussion

During the past decades, extensive studies have evaluated the individual health effect of long-term exposure to $PM_{2.5}$, O_3, and NO_2. However, few studies investigated the joint effects of pollutant mixture in terms of mortality. In this study, we assessed the individual and joint effects between three pollutants and all-causes mortality among 1,406,185 Medicare beneficiaries aged 65 years or older in the southeastern US. The results of Pearson's correlation for annual $PM_{2.5}$, O_3, and NO_2 exposure among 15 southeastern cities suggested

that the concentrations of three pollutants were strongly correlated with each other. Thus, the traditional regression model might not converge or produce an imprecise effect estimates due to collinearity (Bellavia et al., 2019). In our analysis, to investigate the potential nonlinear and non-additive relationship between pollutants and all-cause mortality as well as identify potential interaction between pollutants, we applied the BKMR method to estimate the effects of $PM_{2.5}$, O_3, and NO_2 (Bobb et al., 2015). To the best of our knowledge, this is the first study to use the BKMR method to evaluate the joint effect of multi-pollutant on all-cause mortality.

Using the BKMR model, we found a positive joint effect of overall pollutant mixtures on all-cause mortality. In particular, the joint associations of three pollutants were significantly positive when three pollutants at or above their 65^{th} percentile, as compared to all three pollutants at their median values (Figure 1). Further, the PIP values indicated that the overall pollutant mixture was strongly associated with all-cause mortality. In addition, each pollutant displayed a strong effect with the PIP value higher than 0.5 most of time (PIP value larger than 0.5 is plausibly an important predictor of outcome). Among the three pollutants, $PM_{2.5}$ had the greatest proportion of PIP values that were always larger than the threshold, indicating a stronger association with mortality as compared with O_3 and NO_2. In the single-pollutant analysis, we found that $PM_{2.5}$ was the only pollutant that presented a significantly positive effect on all-cause mortality and its association increased when other pollutants were at their lower quartile. In contrast, the relationship of O_3 and NO_2 with all-cause mortality is not statistically significant. The results were consistent with the conclusions yielded from the univariate C-R function, where we observed a significantly increasing and

nonlinear relationship for $PM_{2.5}$ as well as non-significant relationships for O_3 and NO_2. It is worth noting that the pattern of association in the overall pollutant mixture effect bears strong resemblance with that of the $PM_{2.5}$. This might be explained by the PIP values in which $PM_{2.5}$ was identified as the most important contributor among the three pollutants. Another plausible reason is that we observed these patterns by chance since the variability of both whole-mixture and $PM_{2.5}$ was high at low and high concentrations. Overall, the BKMR results provide evidence that the effects of $PM_{2.5}$ dominate the overall joint effects of pollutant mixtures, especially when at their low concentrations.

Our finding suggested a significantly positive association between $PM_{2.5}$ and all-cause mortality, which was consistent with findings in previous studies focusing on single ambient air pollutants. For example, a study (Franklin et al., 2007) in 27 US communities reported a 1.21% (95% CI: 0.29, 2.14%) increase for all-cause mortality with a 10 $\mu g/m^3$ increase in $PM_{2.5}$; A Chinese prospective cohort study (Li et al., 2018) also indicated a positive effect of $PM_{2.5}$ (HR=1.08; 95% CI: 1.06–1.09) on all-cause mortality among adults aged 65 years and older in China. However, no significant association was observed for NO_2 in our study. Despite several epidemiological studies that reported a positive relationship between NO_2 and all-causes mortality (Lipsett et al., 2011; Cesaroni et al., 2013; Beelen et al., 2014), it is worth noting that they only focused on a single pollutant without disentangling the effects of NO_2 from other pollutants. This limitation was highlighted by a recent review conducted by the WHO REVIHAAP project (WHO, 2013). The REVIHAAP project assessed the emerging evidence on the health effect of NO_2 and concluded that it is difficult to evaluate the individual effects of NO_2 since NO_2 is often highly correlated with other pollutants.

Consequently, the inconsistency between the results in our study and previous studies was likely because we account for the potential nonlinear and non-additive C-R relationship in our model.

The association between O_3 and all-cause mortality was also insignificant in our study. The existing evidence for association between O_3 and mortality is mixed. A cohort study (Jerrett et al., 2009) with 448,850 subjects indicated that long-term exposure to ozone was not associated with all-cause mortality, while a meta-analysis (Huangfu and Atkinson, 2020) incorporating 20 studies for O_3 reported a weak association for peak O_3 and mortality. Noting these studies are unable to take the collinearity problem into account, our study therefore potentially provides a more valid result to disclose the relationship between O_3 and all-cause mortality (Hong et al., 1999).

To explore possible interactions, we assessed the bivariate C-R function and found there are non-interactive effects among three ambient air pollutants. This result is consistent with research conducted in China using BKMR to investigate the association between pollutant mixture and cardiovascular disease (Tong et al., 2018). Notably, we observed that the C-R curve of NO_2 presented a non-parallel trend at high concentration of NO_2, presenting a steeper slope when the concentration of $PM_{2.5}$ at 25^{th} percentile. Since no prior studies have explored the potential interaction of air pollutant mixture on all-cause mortality, we hypothesized that this lack of statistically significant interaction might be due to the relatively wide credible intervals. The sub-population exposed to high level air pollution were likely to be small and weaken the study power, which would distort the C-R trend. As

shown in Figure 5, the 95% CIs were quite wide at high concentration levels. Thus, further research on air pollutant mixtures is needed to explore the potential interactive effect in pollutant mixtures and especially focus on where there are higher numbers of people co-exposed to high levels of co-pollutants. The innovative BKMR approach can also be applicable to explore the joint and individual effects of air pollutants and other risk factors (e.g. temperature, and physical activity).

Our study has several strengths. Firstly, we employed a flexible statistical method, BKMR, to evaluate the joint and individual effects of the pollutant mixtures and visualize the potential nonlinear C-R relationships and interactions among pollutants. Secondly, we used PIP to rank the importance of each pollutant and identify the "bad actor" Thirdly, this study assessed ambient pollutant concentrations using a well-validated ensemble machine learning model that could provide finer-resolution exposure estimates to reduce the potential measurement error.

This study also has several limitations. Firstly, although the BKMR model can capture the nonlinear and non-additive C-R relationship with other pollutants fixed at a certain level, the results of marginal effect can be biased if co-exposures are highly correlated (Zhang et al., 2019). Secondly, the applicability of BKMR to large scale datasets is limited since it requires $O(n^3)$ in time. Therefore, our 15-city study has limited generalizability, particularly given that the air pollution profiling and demographic characters vary much across the US. Thirdly, we cannot rule out of the possibility of unmeasured confounding. For example, the co-

existing chronic diseases that are associated with both air pollution and further death in elderly are not available in the Medicare enrollment file.

5. Conclusion

Using BKMR, we assessed the individual and joint effects of long-term exposure to $PM_{2.5}$, O_3 and NO_2 on all-cause mortality in an elderly cohort. Our results suggest a significantly positive association between pollutant mixture and all-cause mortality, which was mainly driven by $PM_{2.5}$. We found no interaction among the three pollutants. Due to the lack of statistical significance, further studies are needed to investigate where there really is no interaction relationship among the three pollutants.

Acknowledgements

This study was supported by the HERCULES Center P30ES019776 and the National Institute of Environmental Health Sciences (NIEHS R21 ES032606). The authors acknowledge Dr. Joel Schwartz's lab for the preparation of the estimated air pollution data.

References

[1] Allen, M.P., 1997. The problem of multicollinearity. Understanding regression analysis, 176-180.

[2] Ashrap, P., Watkins, D.J., Mukherjee, B., Boss, J., Richards, M.J., Rosario, Z., Vélez-Vega, C.M., Alshawabkeh, A., Cordero, J.F., Meeker, J.D., 2020. Maternal blood metal and metalloid concentrations in association with birth outcomes in Northern Puerto Rico. Environment international 138, 105606.

[3] Barzeghar, V., Sarbakhsh, P., Hassanvand, M.S., Faridi, S., Gholampour, A., 2020. Long-term trend of ambient air PM10, PM2. 5, and O3 and their health effects in Tabriz city, Iran, during 2006–2017. Sustainable Cities and Society 54, 101988.

[4] Beelen, R., Raaschou-Nielsen, O., Stafoggia, M., Andersen, Z.J., Weinmayr, G., Hoffmann, B., Wolf, K., Samoli, E., Fischer, P., Nieuwenhuijsen, M., 2014. Effects of long-term exposure to air pollution on natural-cause mortality: an analysis of 22 European cohorts within the multicentre ESCAPE project. The lancet 383, 785-795.

[5] Bellavia, A., James-Todd, T., Williams, P.L., 2019. Approaches for incorporating environmental mixtures as mediators in mediation analysis. Environment international 123, 368-374.

[6] Bobb, J.F., Henn, B.C., Valeri, L., Coull, B.A., 2018. Statistical software for analyzing the health effects of multiple concurrent exposures via Bayesian kernel machine regression. Environmental Health 17, 1-10.

[7] Bobb, J.F., Valeri, L., Claus Henn, B., Christiani, D.C., Wright, R.O., Mazumdar, M., Godleski, J.J., Coull, B.A., 2015. Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. Biostatistics 16, 493-508.

[8] Brook, J.R., Burnett, R.T., Dann, T.F., Cakmak, S., Goldberg, M.S., Fan, X., Wheeler, A.J., 2007. Further interpretation of the acute effect of nitrogen dioxide observed in Canadian time-series studies. Journal of exposure science & environmental epidemiology 17, S36-S44.

[9] Burnett, R., Chen, H., Szyszkowicz, M., Fann, N., Hubbell, B., Pope, C.A., Apte, J.S., Brauer, M., Cohen, A., Weichenthal, S., 2018. Global estimates of mortality associated with long-term exposure to outdoor fine particulate matter. Proceedings of the National Academy of Sciences 115, 9592-9597.

[10] Cesaroni, G., Badaloni, C., Gariazzo, C., Stafoggia, M., Sozzi, R., Davoli, M., Forastiere, F., 2013. Long-term exposure to urban air pollution and mortality in a cohort of more than a million adults in Rome. Environmental health perspectives 121, 324-331.

[11] Chen, R., Pan, G., Kan, H., Tan, J., Song, W., Wu, Z., Xu, X., Xu, Q., Jiang, C., Chen, B., 2010. Ambient air pollution and daily mortality in Anshan, China: a time-stratified case-crossover analysis. Science of the total environment 408, 6086-6091.

[12] Cohen, A.J., Brauer, M., Burnett, R., Anderson, H.R., Frostad, J., Estep, K., Balakrishnan, K., Brunekreef, B., Dandona, L., Dandona, R., 2017. Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015. The Lancet 389, 1907-1918.

[13] Crouse, D.L., Peters, P.A., Hystad, P., Brook, J.R., van Donkelaar, A., Martin, R.V., Villeneuve, P.J., Jerrett, M., Goldberg, M.S., Pope III, C.A., 2015. 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). Environmental health perspectives 123, 1180-1186.

[14] Deng, W., Liu, J.Z., Coull, B., Lake, E., 2018. Cross-Validated Kernel Ensemble: Robust Hypothesis Test for Nonlinear Effect with Gaussian Process. arXiv preprint arXiv:1811.11025.

[15] Di, Q., Amini, H., Shi, L., Kloog, I., Silvern, R., Kelly, J., Sabath, M.B., Choirat, C., Koutrakis, P., Lyapustin, A., 2019a. Assessing NO2 concentration and model uncertainty with high spatiotemporal resolution across the contiguous United States using ensemble model averaging. Environmental science & technology 54, 1372-1384.

[16] Di, Q., Amini, H., Shi, L., Kloog, I., Silvern, R., Kelly, J., Sabath, M.B., Choirat, C., Koutrakis, P., Lyapustin, A., 2019b. An ensemble-based model of PM2. 5 concentration across the contiguous United States with high spatiotemporal resolution. Environment international 130, 104909.

[17] Domingo-Relloso, A., Grau-Perez, M., Briongos-Figuero, L., Gomez-Ariza, J.L., Garcia-Barrera, T., Dueñas-Laita, A., Bobb, J.F., Chaves, F.J., Kioumourtzoglou, M.-A., Navas-Acien, A., 2019. The association of urine metals and metal mixtures with cardiovascular incidence in an adult population from Spain: the Hortega Follow-Up Study. International journal of epidemiology 48, 1839-1849.

[18] Fann, N., Lamson, A.D., Anenberg, S.C., Wesson, K., Risley, D., Hubbell, B.J., 2012. Estimating the national public health burden associated with exposure to ambient PM2. 5 and ozone. Risk Analysis: An International Journal 32, 81-95.

[19] Faustini, A., Rapp, R., Forastiere, F., 2014. Nitrogen dioxide and mortality: review and meta-analysis of long-term studies. European Respiratory Journal 44, 744-753.

[20] Franklin, M., Zeka, A., Schwartz, J., 2007. Association between PM 2.5 and all-cause and specific-cause mortality in 27 US communities. Journal of exposure science & environmental epidemiology 17, 279-287.

[21] Hastie, T., Tibshirani, R., Friedman, J., 2009. The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.

[22] Hong, Y.-C., Leem, J.-H., Ha, E.-H., Christiani, D.C., 1999. PM (10) exposure, gaseous pollutants, and daily mortality in Inchon, South Korea. Environmental health perspectives 107, 873-878.

[23] Huangfu, P., Atkinson, R., 2020. Long-term exposure to NO2 and O3 and all-cause and respiratory mortality: A systematic review and meta-analysis. Environment International 144, 105998.

[24] Jerrett, M., Burnett, R.T., Pope III, C.A., Ito, K., Thurston, G., Krewski, D., Shi, Y., Calle, E., Thun, M., 2009. Long-term ozone exposure and mortality. New England Journal of Medicine 360, 1085-1095.

[25] Li, J., Zhang, X., Li, G., Wang, L., Yin, P., Zhou, M., 2021. Short-term effects of ambient nitrogen dioxide on years of life lost in 48 major Chinese cities, 2013–2017. Chemosphere 263, 127887.

[26] Li, T., Zhang, Y., Wang, J., Xu, D., Yin, Z., Chen, H., Lv, Y., Luo, J., Zeng, Y., Liu, Y., 2018. All-cause mortality risk associated with long-term exposure to ambient PM2· 5 in China: a cohort study. The Lancet Public Health 3, e470-e477.

[27] Lipsett, M.J., Ostro, B.D., Reynolds, P., Goldberg, D., Hertz, A., Jerrett, M., Smith, D.F., Garcia, C., Chang, E.T., Bernstein, L., 2011. Long-term exposure to air pollution and cardiorespiratory disease in the California teachers study cohort. American journal of respiratory and critical care medicine 184, 828-835.

[28] Liu, D., Lin, X., Ghosh, D., 2007. Semiparametric regression of multidimensional genetic pathway data: Least-squares kernel machines and linear mixed models. Biometrics 63, 1079-1088.

[29] Loh, W.Y., 2011. Classification and regression trees. Wiley interdisciplinary reviews: data mining and knowledge discovery 1, 14-23.

[30] Orioli, R., Cremona, G., Ciancarella, L., Solimini, A.G., 2018. Association between PM10, PM2. 5, NO2, O3 and self-reported diabetes in Italy: A cross-sectional, ecological study. PloS one 13, e0191112.

[31] Plass, D., Tobollik, M., Wintermeyer, D., 2019. Burden of disease due to nitrogen dioxide exposure in Germany. European Journal of Public Health 29, ckz185. 657.

[32] Proust-Lima, C., Letenneur, L., Jacqmin-Gadda, H., 2007. A nonlinear latent class model for joint analysis of multivariate longitudinal data and a binary outcome. Statistics in medicine 26, 2229-2245.

[33] Requia, W.J., Di, Q., Silvern, R., Kelly, J.T., Koutrakis, P., Mickley, L.J., Sulprizio, M.P., Amini, H., Shi, L., Schwartz, J., 2020. An ensemble learning approach for estimating high spatiotemporal resolution of ground-level ozone in the contiguous United States. Environmental Science & Technology 54, 11037-11047.

[34] Roberts, S., Martin, M.A., 2006. Investigating the mixture of air pollutants associated with adverse health outcomes. Atmospheric Environment 40, 984-991.

[35] Shi, L., Wu, X., Yazdi, M.D., Braun, D., Awad, Y.A., Wei, Y., Liu, P., Di, Q., Wang, Y., Schwartz, J., 2020. Long-term effects of PM2· 5 on neurological disorders in the American Medicare population: a longitudinal cohort study. The Lancet Planetary Health 4, e557-e565.

[36] Tolbert, P.E., Klein, M., Peel, J.L., Sarnat, S.E., Sarnat, J.A., 2007. Multipollutant modeling issues in a study of ambient air quality and emergency department visits in Atlanta. Journal of exposure science & environmental epidemiology 17, S29-S35.

[37] Tong, Y., Luo, K., Li, R., Pei, L., Li, A., Yang, M., Xu, Q., 2018. Association between multi-pollutant mixtures pollution and daily cardiovascular mortality: An exploration of exposure-response relationship. Atmospheric Environment 186, 136-143.

[38] Vodonos, A., Awad, Y.A., Schwartz, J., 2018. The concentration-response between long-term PM2. 5 exposure and mortality; A meta-regression approach. Environmental research 166, 677-689.

[39] Wang, Y., Ying, Q., Hu, J., Zhang, H., 2014. Spatial and temporal variations of six criteria air pollutants in 31 provincial capital cities in China during 2013–2014. Environment international 73, 413-422.

[40] Wei, Y., Wang, Y., Wu, X., Di, Q., Shi, L., Koutrakis, P., Zanobetti, A., Dominici, F., Schwartz, J.D., 2020. Causal effects of air pollution on mortality rate in Massachusetts. American journal of epidemiology 189, 1316-1323.

[41] World Health Organization, 2013. Review of evidence on health aspects of air pollution–REVIHAAP project: final technical report. Bonn: WHO European Centre for Environment and Health.

[42] World Health Organization, 2016. Ambient air pollution: A global assessment of exposure and burden of disease.
[43] Yin, S., Wang, C., Wei, J., Wang, D., Jin, L., Liu, J., Wang, L., Li, Z., Ren, A., Yin, C., 2020. Essential trace elements in placental tissue and risk for fetal neural tube defects. Environment international 139, 105688.
[44] Yu, W., Guo, Y., Shi, L., Li, S., 2020. The association between long-term exposure to low-level PM2. 5 and mortality in the state of Queensland, Australia: A modelling study with the difference-in-differences approach. PLoS medicine 17, e1003141.
[45] Zhang, Y., Dong, T., Hu, W., Wang, X., Xu, B., Lin, Z., Hofer, T., Stefanoff, P., Chen, Y., Wang, X., 2019. Association between exposure to a mixture of phenols, pesticides, and phthalates and obesity: comparison of three statistical models. Environment international 123, 325-336.
[46] Zhao, N., Smargiassi, A., Hudson, M., Fritzler, M.J., Bernatsky, S., 2020. Investigating associations between anti-nuclear antibody positivity and combined long-term exposures to NO2, O3, and PM2. 5 using a Bayesian kernel machine regression approach. Environment international 136, 105472.

Table

Table 1: Cohort characteristics

	Cohort (n=1406185)
Age at entry, years	
Minimum	66
Median	66
Maximum	71
Mean	67(1.4)
Sex	
Men	587762(41.8%)
Women	818423(58.2%)
Race	
White	1066639(75.9%)
Black	299252(21.3%)
Other*	40294(2.9%)
Temperature, $Celsius$	16.6(1.2)
Relative humidity, %	72.8(3.3)
Population density, people per mile^2	1603.8(1186.6)
Percent black, %	29.1(23.7)
Not graduated from high school, %	26.8(13.8)
Below poverty level, %	9.7(6.1)
Median house value, US$1000	163.8(84.7)
Median household income, US$1000	51.5(17.9)
Owner-occupied housing, %	65.2(14.7)
Smoking rate, %	44.1(4.8)
Body-mass index, kg/m^2	27.0(0.5)
PM$_{2.5}$**,** $\mu g/m^3$	13.1(2.0)
NO$_2$**,** ppb	21.6(7.3)
O$_3$**,** ppb	47.3(3.4)

Data are n (%) or mean (SD). *Other included Asian, Hispanic, American Indian or Alaskan Native, and unknown.

Figure Legends

Figure 1. Overall effects of $PM_{2.5}$, O_3, and NO_2 with 95% CI. The figure shows the estimated change in risk of all-cause mortality when three pollutants were set at particular percentiles (ranging from 25^{th} to 75^{th}) compared to when all pollutants are all at their 50^{th} percentile.

Figure 2. (a) The histogram of PIP's for $PM_{2.5}$, O_3, and NO_2; (b) PIP trace plot for $PM_{2.5}$, O_3, and NO_2. In this plot, each colored line represents a PIP trace for a certain variable. "threshold" is the value based on which we determine the variable to be included or not. For instance, at threshold 0.5, we categorize PM2.5 as selected if its PIP is greater than 0.5, and discarded if less than 0.5. y-axis is the proportion of the PIP counts which are greater than a certain threshold. Therefore, as threshold approaches 1, the proportions are decreasing. Moreover, for important variables, we expect their proportions to be as large as possible.

Figure 3. Single-pollutant association with mortality. The plot shows the change in risks of all-cause mortality with an 95% credible inter in a single pollutant, when all other pollutants were fixed at either 25^{th}, 50^{th}, or 75^{th} percentile.

Figure 4. The univariate concentration-response functions with 95% confidence bands (shaded areas) for each pollutant ($PM_{2.5}$, O_3, and NO_2) with the other pollutants fixed at the median.

Figure S1. Graphical display of the pairwise correlation coefficients between $PM_{2.5}$, NO_2, and O_3.

Figure S2. Bivariate exposure response functions for each pollutant presented on x-axis when pollutant on y-axis was fixed at 25% (red line), 50% (green line), and 75% (blue line) percentile respectively, and other pollutants were fixed at their median.

Figures

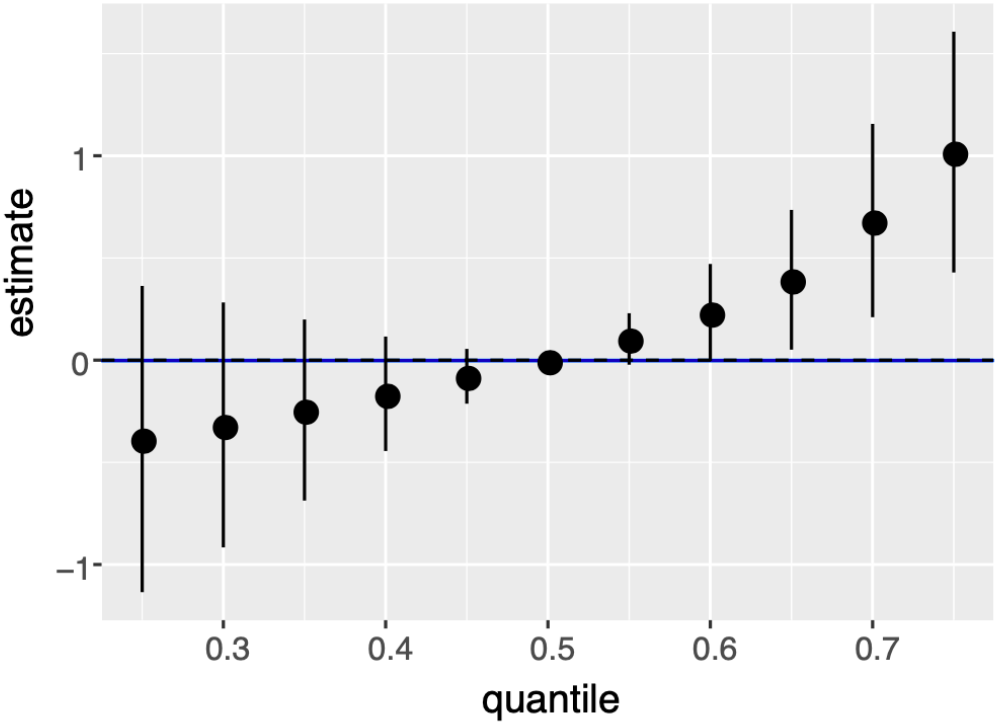

Figure 1

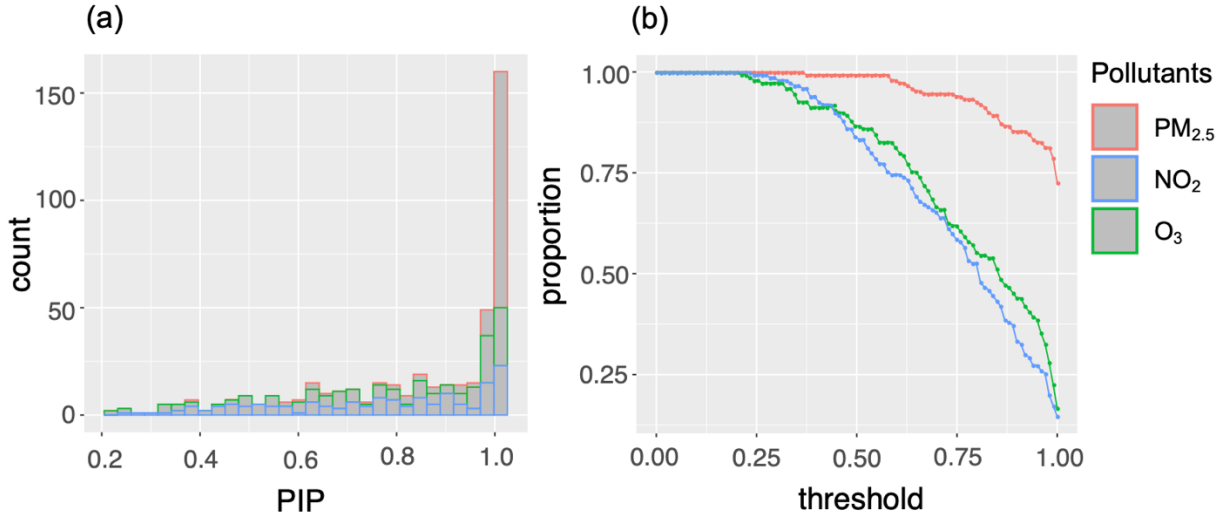

Figure 2

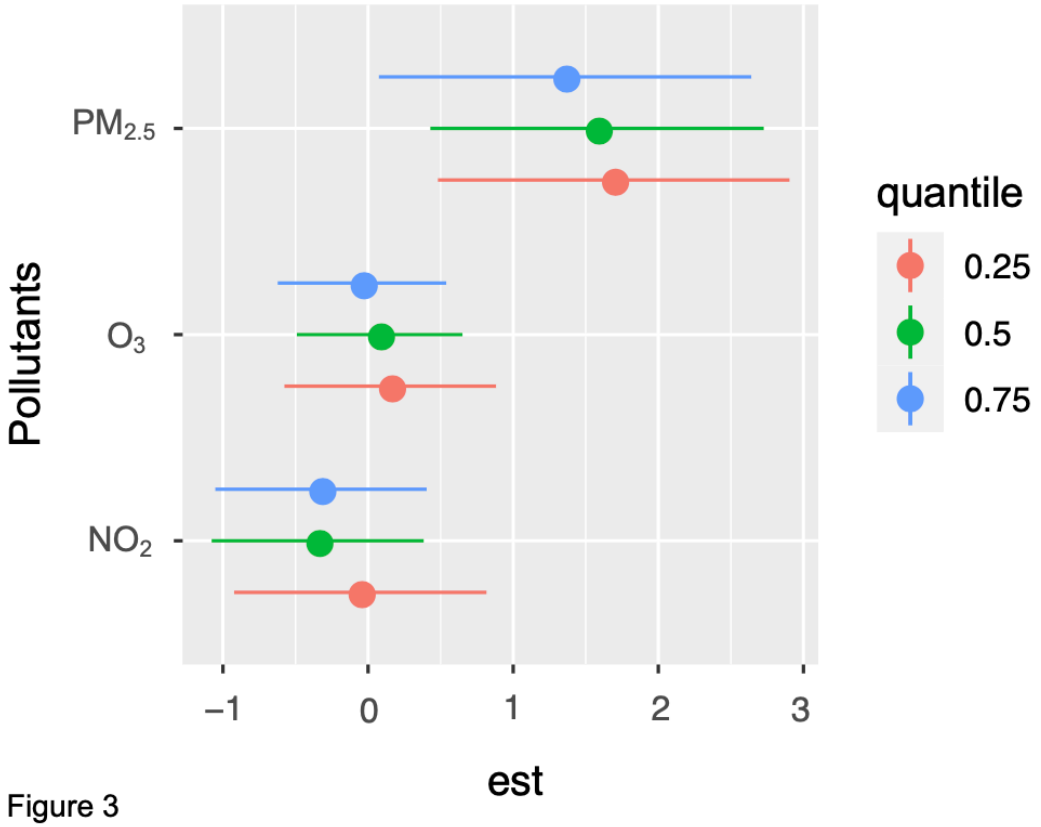

Figure 3

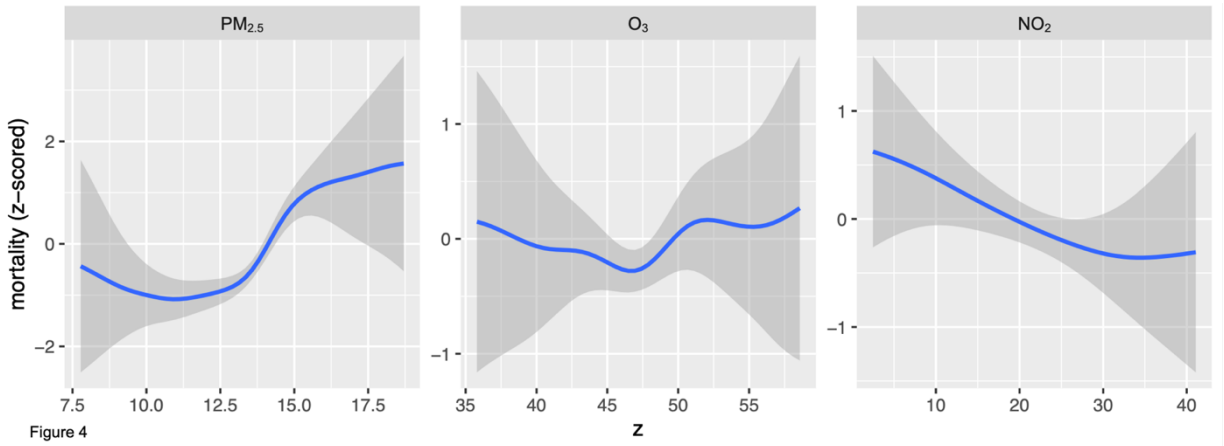

Figure 4

Supplementary Materials

Our treatment of kernel-machine regression and variable selection follows Bobb et al. (2015).

Overview of kernel-machine regression

We assume that the interaction function $h : \mathbb{R}^q \to \mathbb{R}$ lies in a function space $\mathcal{H}_K$ generated by a positive definite kernel function $K : \mathbb{R}^q \times \mathbb{R}^q \to \mathbb{R}$. A kernel function $K(\mathbf{z}, \mathbf{z}')$ takes two arguments: $\mathbf{z} = (z_1, \ldots, z_q)^\top$, which represents the vector of pollutant components for one individual, and $\mathbf{z}' = (z'_1, \ldots, z'_q)^\top$, the vector of pollutant components for a second individual. From Mercer's theorem (Cristianini and Shawe-Taylor, 2000), under some regularity conditions, a kernel function implicitly specifies a unique function space spanned by a particular set of orthogonal basis function (features) $\{\phi_l(\mathbf{z})\}_{l=1}^L$. Therefore, any $h(\mathbf{z}) \in \mathcal{H}_K$ can be represented using some set of coefficients $\{\eta_l\}_{l=1}^L$ as $h(\mathbf{z}) = \sum_{l=1}^{L} \eta_l \phi_l(\mathbf{z})$ (the primal representation). Alternatively, $h(\mathbf{z})$ can also be represented using a kernel function $K(\cdot, \cdot)$ as $h(\mathbf{z}) = \sum_{i=1}^{n} K(\mathbf{z}_i, \mathbf{z})\alpha_i$ for some set of coefficients $\{\alpha_i\}_{i=1}^n$ (the dual representation). For a multidimensional $\mathbf{z}$, it is more convenient to specify $h(\mathbf{z})$ using the dual representation, because explicit basis functions might be complicated to specify and the number of features might be high or even infinite.

Examples of this correspondence include *the 1st polynomial kernel (linear kernel)*: $K(\mathbf{z}, \mathbf{z}') = 1 + z_1 z'_1 + \ldots + z_q z'_q$ with basis functions $\{\phi_l(\mathbf{z})\} = \{z_1, \ldots, z_q\}$; *the 2nd polynomial kernel (quadratic kernel)*: $K(\mathbf{z}, \mathbf{z}') = (1 + z_1 z'_1 + \ldots + z_q z'_q)^2$ with basis functions $\{\phi_l(\mathbf{z})\} = \{z_k, z_k z_{k'}\}(k, k' = 1, \ldots, q)$; and *the Gaussian kernel*:

$K(\mathbf{z}, \mathbf{z}') = \exp\{-\sum_{j=1}^{q} \rho(z_j - z_j')^2\}$, which generates the function space spanned by radial basis functions and ρ is the length-scale parameter. BKMR focuses on the Gaussian kernel, which represents a nonparametric model with a high degree of smoothness (i.e. infinitely differentiable) that can incorporate more general types of nonlinearity. Liu et al. (2007) argued defining $\tau = \sigma^2/\lambda$, Equation (1) with h specified in the dual form can be expressed as a linear mixed model:

$$Y_i \sim N(h(\mathbf{z}_i) + \mathbf{x}_i^\top \boldsymbol{\beta}, \sigma^2), \quad (2)$$

$$\mathbf{h} \equiv (h(\mathbf{z}_1), \ldots, h(\mathbf{z}_n))^\top \sim N(\mathbf{0}, \tau \mathbf{K}),$$

where λ is the regularization parameter, and $\mathbf{K}$ has (i, j) element as $K(\mathbf{z}_i, \mathbf{z}_j)$.

Assessing Variable Importance using Posterior Inclusion Probability (PIP)

We further performed variable selection to provide the posterior inclusion probabilities (PIP) for pollutant components, and variables with a PIP greater than 0.5 are considered as a meaningful predictor (Barbieri and Berger, 2004). To perform variable selection within a Bayesian paradigm using the Gaussian kernel, one can borrow the idea of Automatic Relevance Determination (ARD), which means that the kernel has one length-scale per variable. They define the augmented Gaussian kernel function as $K(\mathbf{z}, \mathbf{z}'; \boldsymbol{\rho}) = \exp\{-\sum_{j=1}^{q} \rho_j (z_j - z_j')^2\}$, where $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_q)^\top$, and $\mathbf{K}_\rho$ to be the $n \times n$ matrix with (i, j) element to be $K(\mathbf{z}_i, \mathbf{z}_j; \boldsymbol{\rho})$. Assuming a "slab-and-spike" prior on the auxiliary parameters $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_q)^\top$,

$$\rho_j | \delta_j \sim \delta_j f_1(\rho_j) + (1 - \delta_j) P_0, \quad (3)$$

$$\delta_j \sim \text{Bernoulli}(\pi),$$

where $f_1(\cdot)$ is a pdf with support on $\mathbb{R}^+$ and P_0 is the density with point mass at 0. The posterior mean of the indicator δ_j has the natural interpretation as the posterior probability that component j is an important component of the mixture, or the posterior inclusion probability (PIP) of component j.

Supplementary Figures

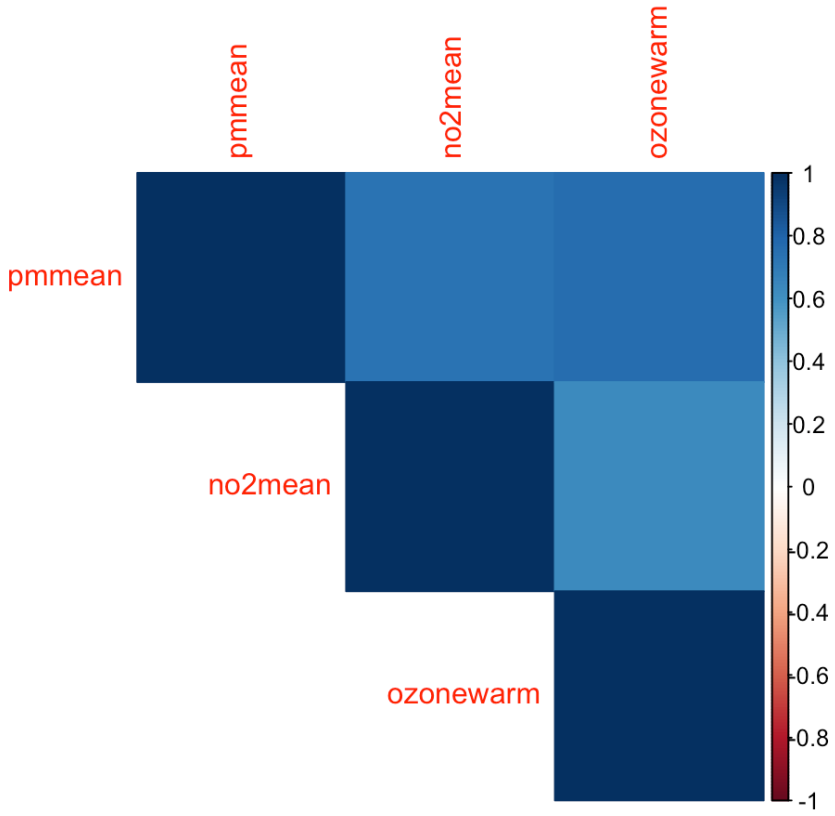

Figure S1. Graphical display of the pairwise correlation coefficients between $PM_{2.5}$, NO_2, and O_3.

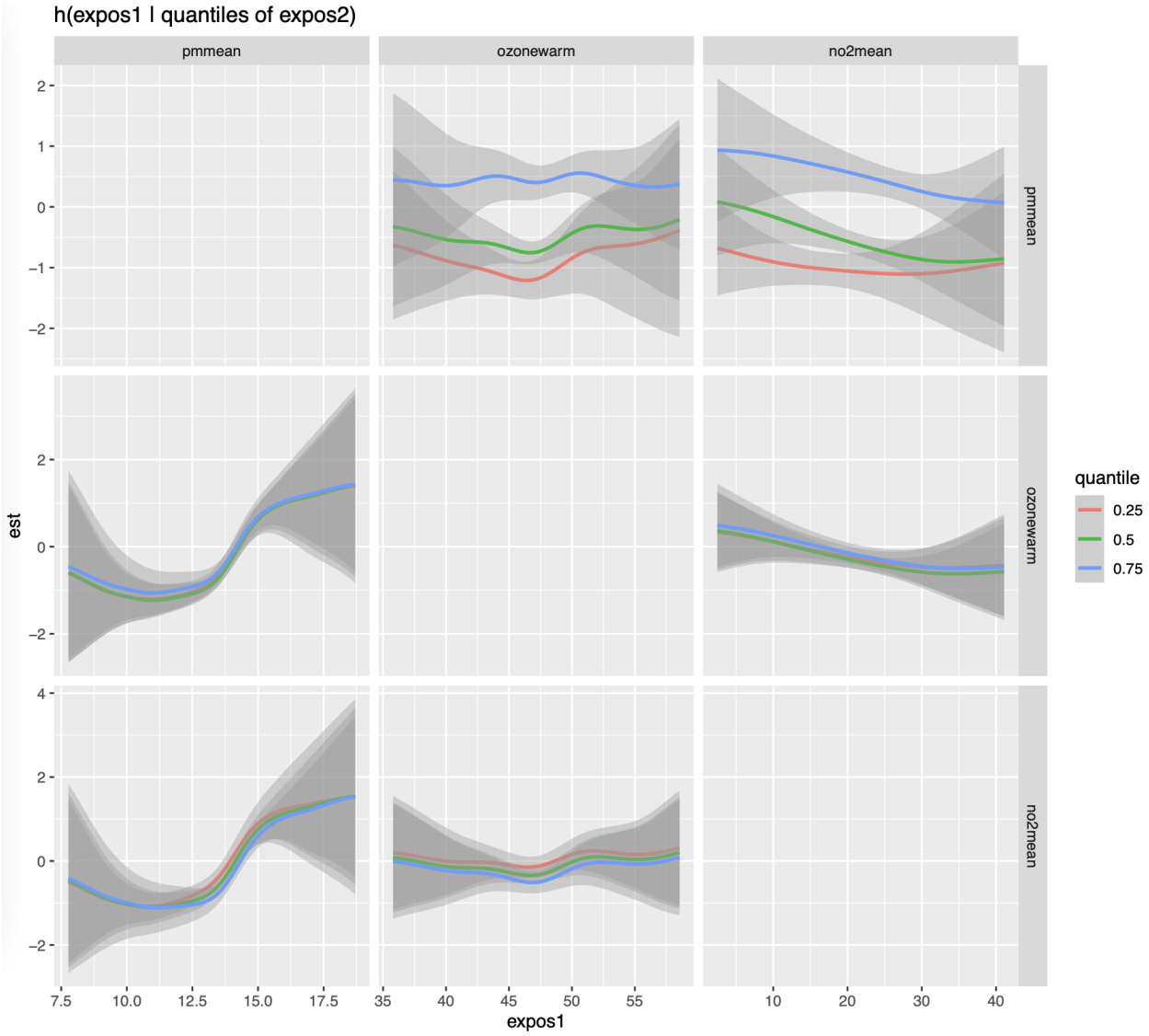

Figure S2. Bivariate exposure response functions for each pollutant presented on x-axis when pollutant on y-axis was fixed at 25% (red line), 50% (green line), and 75% (blue line) percentile respectively, and other pollutants were fixed at their median.