HHS Public Access

Author manuscript

Epidemiology. Author manuscript; available in PMC 2023 March 01.

Published in final edited form as:

Epidemiology. 2022 March 01; 33(2): 157–166. doi:10.1097/EDE.0000000000001447.

Estimating associations between annual concentrations of particulate matter and mortality in the US, using data linkage and Bayesian Maximum Entropy

Jacqueline E. Rudolph[1], **Stephen R. Cole**[2], **Jessie K. Edwards**[2], **Eric A. Whitsel**[2,3], **Marc L. Serre**[4], **David B. Richardson**[2]

[1]Department of Epidemiology, Johns Hopkins University

[2]Department of Epidemiology, University of North Carolina at Chapel Hill

[3]Department of Medicine, University of North Carolina at Chapel Hill

[4]Department of Environmental Sciences and Engineering, University of North Carolina at Chapel Hill

Abstract

Background.—Exposure to fine particulate matter ($PM_{2.5}$) is an established risk factor for human mortality. However, previous US studies have been limited to select cities or regions or to population subsets (e.g., older adults).

Methods.—Here, we demonstrate how to use the novel geostatistical method Bayesian Maximum Entropy to obtain estimates of $PM_{2.5}$ concentrations in all contiguous US counties, 2000–2016. We then demonstrate how one could use these estimates in a traditional epidemiologic analysis examining the association between $PM_{2.5}$ and rates of all-cause, cardiovascular, respiratory, and (as a negative control outcome) accidental mortality.

Results.—We estimated that, for a 1 log ($\mu g/m^3$) increase in $PM_{2.5}$ concentration, the conditional all-cause mortality incidence rate ratio (IRR) was 1.029 (95% CI: 1.006, 1.053). This implies that the rate of all-cause mortality at 10 $\mu g/m^3$ would be 1.020 times the rate at 5 $\mu g/m^3$. IRRs were larger for cardiovascular mortality than for all-cause mortality in all gender and race–ethnicity groups. We observed larger IRRs for all-cause, non-accidental, and respiratory mortality in Black non-Hispanic Americans than White non-Hispanic Americans. However, our negative control analysis indicated the possibility for unmeasured confounding.

Conclusions.—We used a novel method that allowed us to estimate $PM_{2.5}$ concentrations in all contiguous US counties and obtained estimates of the association between $PM_{2.5}$ and mortality comparable to previous studies. Our analysis provides one example of how Bayesian Maximum

Correspondence: Jacqueline E. Rudolph, PhD, Department of Epidemiology, Bloomberg School of Public Health, Johns Hopkins University, 615 N. Wolfe St., Baltimore, MD 21205, (410) 502-9048, jacqueline.e.rudolph@emory.edu.

Conflicts of Interest: The authors have no conflicts to declare.

Data and Code: All data were obtained from national databases. Please visit the following websites: https://www.cdc.gov/nchs/data_access/cmf.htm (mortality data), https://www.epa.gov/aqs (air pollution data), and data.census.gov (US Census and American Community Survey data). Code for this analysis can be made available upon request.

Entropy could be used in epidemiologic analyses; future work could explore other ways to use this approach to inform important public health questions.

Keywords

particulate matter; mortality; spatial interpolation; Bayesian Maximum Entropy

Introduction

As one of the six criteria air pollutants, particulate matter (PM) is monitored and regulated by the United States (US) Environmental Protection Agency (EPA). National Ambient Air Quality Standards have been set for ambient concentrations of PM. In part due to these regulations, ambient concentrations of PM have steadily dropped in recent years.[1–3] The EPA estimates that between 2000 and 2015 daily concentrations of PM with an aerodynamic diameter <10 μm (PM_{10}) and PM with an aerodynamic diameter <2.5μm ($PM_{2.5}$) improved by 34% and 40%, respectively.[4] Even so, PM remains a ubiquitous exposure, and a large body of evidence has established its adverse environmental and health effects, even at concentrations below the current regulatory action level.[1,3,5,6]

Of all the sub-classifications of PM, $PM_{2.5}$ has been most consistently associated with all-cause, cardiovascular, and respiratory mortality following acute and long-term exposure.[7–18] The EPA determined in the 2009 and 2019 Integrated Science Assessments that $PM_{2.5}$ has a causal effect on mortality and cardiovascular outcomes (and is “likely to be causal” for respiratory outcomes).[1,3] One limitation of past research on the association between $PM_{2.5}$ and mortality in the US, though, has been the restricted target populations. Often, the restrictions have been geographic, i.e. limited to a single city or state, a select group of cities, or a subset of US counties.[7,13,16,17,19–21] This is understandable if the goal is to study unique, localized events, to reduce bias due to spatial confounding, or to investigate research questions within a more geographically limited target population. However, researchers and policymakers might also be interested in larger target populations, such as the entire US population.

Here, we demonstrate how the advanced geostatistical method Bayesian Maximum Entropy can be used to interpolate $PM_{2.5}$ monitor data to inform the annual concentrations in all counties in the contiguous US, even those that lack monitors. We then demonstrate how one might use these exposure estimates in an analysis assessing the association between $PM_{2.5}$ and county-level mortality rates, by fusing the exposure estimates to data from the National Center for Health Statistics and the US Census.

METHODS

Data sources

We obtained annual arithmetic average $PM_{2.5}$ concentrations in $\mu g/m^3$ from the EPA’s Air Quality Systems Database, which includes data on $PM_{2.5}$ concentrations measured at EPA monitoring sites across the US since 1999. While the network of EPA monitors for $PM_{2.5}$ is well dispersed across the country (see eFigure 1), their placement is based on the population

size and past air quality within a given geographical area. The location is further intended to reflect regional transport of ambient $PM_{2.5}$.[1] The exact location of monitors changed over time, and there was not a monitor in every US county in every year. Thus, we interpolated the monitor data to obtain county-specific concentrations, as described below.

We obtained mortality counts and mid-year population counts from National Center for Health Statistics (NCHS) data aggregated by year, county, age-category (<1 year, 1–4, 5–9, 10–14, 15–19, 20–24, 25–34, 35–44, 45–54, 55–64, 65–74, 75–84, 85+), gender (male or female), race (White, Black, American Indian or Alaskan Native, Asian or Pacific Islander), and Hispanic ethnicity (yes, no, or undeclared).[22] Underlying cause of death was coded using International Classification of Diseases Tenth Revision (ICD-10). We defined four categories of mortality that were of substantive interest in our study: all-cause, non-accidental (excluding ICD-10 codes with the prefix S, V, Z, or U), select cardiovascular (ICD-10: I10-I70) and non-cancerous respiratory mortality (ICD-10: J00-J99).

We were additionally interested in accidental mortality (those causes excluded from non-accidental mortality) as a negative control outcome.[9,11,23–25] Negative control outcomes are events that are expected, on the basis of substantive knowledge, to have no causal relationship with the exposure under investigation. In an ideal setting, the estimated association between the exposure and negative control outcome would be null; therefore, a non-null association between these variables can be an indicator of uncontrolled bias in the study, assuming the bias is affecting this association in the same manner as for the outcome of interest. For example, the approach assumes that the unmeasured confounders for the association between the exposure and outcome of interest are also the only unmeasured confounders of the association between the exposure and negative control outcome. We use the findings from the negative control outcome to infer the possible direction but not the magnitude of the bias.[23]

We derived county-level covariates from the 2000 and 2010 US Censuses and from the 2005–2016 American Community Surveys (ACS). For the ACS data, we obtained annual estimates for those counties that had a population greater than 65,000 and 5-year-summary estimates for all counties.[26] Due to the gap between the 2000 Census and the start of the ACS in 2005, we interpolated values for each covariate using natural cubic splines, such that we had one value per county per year.[27] The variables included as potential confounders were those we hypothesized affect county-level, ambient concentrations of $PM_{2.5}$ and mortality. Accordingly, we obtained data on median household income (in 2000-inflation-adjusted dollars), percent of individuals who graduated high school, proportion living in an urban area, and proportion of the county population that was Black. We categorized urbanicity into mostly urban (at least half the population living in an urban area), mostly rural (less than half living in an urban area), and rural (no one living in an urban area). This covariate set was comparable to other ecologic $PM_{2.5}$ studies.[16,17]

Exposure model

To interpolate concentrations of $PM_{2.5}$, we used the Bayesian Maximum Entropy geostatistical method. This estimation approach allowed us to leverage the concept of a space/time random field, $Z(\boldsymbol{p})$, to describe the randomness and correlation in $PM_{2.5}$

measured at space/time points $\boldsymbol{p} = (\boldsymbol{s}, t)$, where $\boldsymbol{s}$ are the spatial coordinates and t is time in calendar year. Bayesian Maximum Entropy has been described in detail elsewhere;[28–30] we summarize below the essential points for our implementation.

Bayesian Maximum Entropy integrates general knowledge and site-specific knowledge about a space/time field. General knowledge can include the mean trend (e.g., that concentrations decreased over time) and covariance between points, which are modeled from the data. This information is used to build the prior probability density function that, outside of what is specified by the general knowledge, weakly constrains the data (i.e., maximizes information entropy).[31,32] Site-specific knowledge is the measured values, i.e. annual average $PM_{2.5}$ concentration at each space/time point. The site-specific knowledge is integrated with the prior probability density function using a Bayesian epistemic knowledge integration rule,[29,33] and a posterior probability density function is obtained at each estimation point.

For our implementation, the model inputs were the annual average $PM_{2.5}$ concentrations at EPA monitors. Our estimation grid was defined by the spatial coordinates of the county population centers (defined in 2015) at which we sought to estimate annual average $PM_{2.5}$ concentrations. We restricted our estimation to counties within the contiguous US.

First, we log transformed the data, such that $Y(\boldsymbol{p}) = \log(Z(\boldsymbol{p}))$. We then obtained the global offset removed field $X(\boldsymbol{p}) = Y(\boldsymbol{p}) - m(\boldsymbol{s}, t)$, so that our estimation was conducted in a field that was more stable in space and time. The global offset was modeled using the formula $m_Y(\boldsymbol{s}, t) = m_s(\boldsymbol{s}) + m_t(t)$, where $m_s(\boldsymbol{s})$ was the smoothed spatial mean component (calculated by applying an exponential spatial filter to the mean $m_t j$ at each time point j). These two quantities were calculated as follows:

$$m_s(\boldsymbol{s}) = \frac{\sum_i \left(ms_i \times \exp\left(-\frac{d_{si}}{S_{range}} \right) \right)}{\sum_i \left(\exp\left(-\frac{d_{si}}{S_{range}} \right) \right)}$$

$$m_t(t) = \frac{\sum_j \left(mt_j \times \exp\left(-\frac{d_{tj}}{T_{range}} \right) \right)}{\sum_j \left(\exp\left(-\frac{d_{tj}}{T_{range}} \right) \right)} - m$$

where $\boldsymbol{s}$ is a location of interest, i is the monitoring station, ms_i is the mean of the measured log $PM_{2.5}$ concentrations at i, d_{si} is the distance between $\boldsymbol{s}$ and i, t is the time of interest, j is the time of the measurement, mt_j is the mean of the measured log $PM_{2.5}$ concentrations measured at time j, and d_{tj} is the time difference between t and j. In the second equation, the average m of all measured log $PM_{2.5}$ concentrations was subtracted to avoid double counting when $m_s(\boldsymbol{s})$ and $m_t(\boldsymbol{t})$ were added. Finally, the spatial range (S_{range}) and temporal range (T_{range}) governed the level of spatial and temporal smoothing, respectively. We chose to calculate a regionally defined global offset (rather than calculating a single mean for the entire US); consequently, we chose a S_{range} of 50 km. The specified T_{range} was 3 years.

Since we subtracted out the global offset, the mean trend for $X(\boldsymbol{p})$ was set to zero, and we found the following nested exponential covariance model below to be a good fit for our data:

$$c_X(r,\tau) = 0.0750\left[\log\left(\mu g/m^3\right)\right]^2\left[0.7\exp\left(-\frac{3r}{50km}\right)\exp\left(-\frac{3\tau}{14year}\right) + 0.3\exp\left(-\frac{3r}{600km}\right)\exp\left(-\frac{3\tau}{1year}\right)\right],$$

where r is the spatial distance and τ the time distance between $\boldsymbol{p}$ and $\boldsymbol{p}'$. To interpret the meaning of the chosen model parameters, we specified that the variance of $X(\boldsymbol{p})$ was 0.075 (log ($\mu g/m^3$))^2, and that 70% of the covariance in the field was explained by variation over a spatial range of 50 km and a temporal range of 14 years (i.e., city-level $PM_{2.5}$ which changes only over long periods of time). Then, 30% of the covariance was explained by variation at a spatial range of 600 km and a temporal range of 1 year (i.e., region-level $PM_{2.5}$ which could change more quickly). Using a nested covariance model with two terms allowed us to account for variation in $PM_{2.5}$ over city- and region-level distances and over short and long periods of time. One of the advantages of space/time BME over other spatial models is this ability to flexibly model space/time correlation.

Lastly, we chose to use a "soft" (i.e., probabilistic) data approach to account for the fact that measurements were not taken every day of the year.[34–36] We denote n_{it} as the number of measurements at monitor i in year t. We treated the annual average from any monitor that recorded measurements at least every 3 days as "hard," non-probabilistic data ($n_{it} \geq 122$). This cut-point was chosen because it reflects the federal reference method for $PM_{2.5}$. Otherwise, data were treated as soft. Specifically, we assumed that the monitor data were normally distributed with mean equal to the recorded arithmetic average (μ_{it}) and a standard deviation specified as follows:

$$\sigma_{it} = \frac{s_{it}}{\sqrt{n_{it}}} \times \sqrt{\frac{365 - n_{it}}{365}},$$

where s_{it} was the recorded arithmetic standard deviation.

Outcome model

Our analytic data set contained the interpolated $PM_{2.5}$ concentrations and sociodemographic covariate data merged to the mortality and population counts by year and county. Using these data, we estimated the covariate-conditional incidence rate ratios (IRR) comparing rates of a given mortality category for a one-unit increase in log $PM_{2.5}$ concentration. To estimate these rate ratios, we used the following Poisson regression model:

$$\log(d) = \beta_0 + \beta_1 \log(pm) + \sum_{k=2}^{40} \beta_k V_k + \log(n),$$

where d is the number of deaths, n is the mid-year population count, and $\boldsymbol{V}$ is the set of covariates, including indicator terms for age category, calendar year, and urbanicity as well as restricted quadratic splines terms (with 3 knots) for income, education, and proportion Black. We obtained 95% confidence intervals (CI) using robust standard errors for the IRR.[37]

We also carried out several supplementary analyses. First, we ran models where the data were stratified by gender and race–ethnicity, effectively including interaction terms between these potential effect measure modifiers and all variables in the model. Second, in contrast to the main analysis which used exposure and outcome data from the same year, we also ran models where there was a lag of 1–5 years between the mortality data and the exposure and confounder data. For example, in the 1-year-lag analysis, we regressed a county's annual rate of mortality against the $PM_{2.5}$ concentration from the previous year. In these lagged analyses, fewer calendar years of data were included because, for example, we lacked exposure data from 5 years before the deaths that occurred in 2000. Finally, we repeated our analysis within the subset of counties that had a monitor within its borders in a given year, to assess the possible impact of exposure misclassification resulting from including counties that did not have a monitor (and whose exposure relied more heavily on the interpolation model).

Data cleaning and outcome modeling were carried out using SAS version 9.4 (SAS Institute, Cary, NC). Exposure modeling was carried out using the BMELIB version 2.0c library for MATLAB (The MathWorks, Inc., Natick, MA).

RESULTS

Across the years 2000–2016, annual average concentrations of $PM_{2.5}$ decreased (Figure 1). The national average $PM_{2.5}$ concentration (using our interpolated estimates) was 12.2 $\mu g/m^3$ in 2000 but 8.0 $\mu g/m^3$ in 2016. The maximum concentration also decreased, from 24.3 $\mu g/m^3$ to 14.4 $\mu g/m^3$ in 2015. The trend in $PM_{2.5}$ concentrations is reinforced by the maps in eFigure 1, which show that the concentrations of measured $PM_{2.5}$ decreased across the entire US. However, high concentrations remained in select areas, such as southern California. eFigure 2 illustrates an example of the model output (interpolated estimates and their variances) from our implementation of Bayesian Maximum Entropy.

Important socioeconomic and demographic characteristics of the US population across our study period are summarized in Table 1. The proportion of the US population who reported being Black or Other race increased from 2000 to 2016 (respectively, from 13.1% to 14.1% and from 4.9% to 7.5%). The proportion who reported being Hispanic or Latinx also increased, from 12.6% to 17.8%. Furthermore, by 2016 more counties were listed as "mostly urban" (i.e., more than half their population living in an urban area) than in 2000, and a higher percent had graduated high school.

Rates of mortality generally decreased between 2000 and 2016 (Table 2). Comparing rates that had been standardized to the age distribution of 2000, we saw that the all-cause mortality rate in 2000 was 865.5 per 100,000 person–years and the rate in 2016 was 736.1 per 100,000 person–years. The rate of cardiovascular mortality in 2000 was 335.7 per 100,000 person–years, compared to 218.7 in 2016. Respiratory mortality rates were 83.1 per 100,000 person–years and 70.1 per 100,000 person–years, respectively, for 2000 and 2016.

As illustrated in Figure 2, we estimated that, for every 1 log ($\mu g/m^3$) increase in $PM_{2.5}$ concentration, the all-cause mortality IRR was 1.029 (95% CI: 1.006, 1.053), conditional

on all covariates in the model. This implies that the rate of all-cause mortality at 10 μg/m^3 would be 1.020 times the rate at 5 μg/m^3 and that the rate at 15 μg/m^3 would be 1.012 times the rate at 10 μg/m^3. Rates of mortality increased as $PM_{2.5}$ concentration increased, with the association being further from the null at lower concentrations than at higher concentrations. Results for non-accidental mortality and respiratory mortality were similar in magnitude to the all-cause mortality IRRs. For cardiovascular mortality, we estimated that the conditional IRR for a 1 log ($\mu g/m^3$) increase in log $PM_{2.5}$ concentration was 1.105 (95% CI: 1.055, 1.157). Finally, we estimated that the IRR for accidental mortality (our negative control outcome) was 0.804 (95% CI: 0.690, 0.938). IRRs became closer to the null as the lag between exposure and outcome increased (Table 3). With a lag of 5 years, the IRR was 1.009 (95% CI: 0.982, 1.037) for all-cause mortality and 1.076 (95% CI: 1.033, 1.122) for cardiovascular mortality. This was also the case for accidental mortality, which had an IRR of 0.821 (95% CI: 0.696, 0.969) when the exposure was lagged by 5 years.

The estimated IRRs stratified by gender were similar to the overall estimates, although men had marginally stronger associations than women (Figure 2 and Table 3). For example, with no lag, we estimated an IRR for cardiovascular mortality in men of 1.115 (95% CI: 1.062, 1.171), compared to an IRR of 1.102 (95% CI: 1.053, 1.154) in women. The results stratified by race-ethnicity were more striking (Figure 3 and Table 4). The IRRs for all-cause, non-accidental, and respiratory mortality among Black non-Hispanic Americans were marginally further from the null than among White non-Hispanic Americans (except with a lag of 5 years). With a lag of 0 years, the IRR for all-cause mortality was 1.087 (95% CI: 1.021, 1.157) for Black non-Hispanic Americans versus 1.047 (95% CI: 1.016, 1.078) for White non-Hispanic Americans. The difference for respiratory mortality was even larger: IRRs of 1.183 (95% CI: 1.066, 1.313) and 1.017 (95%: 0.981, 1.054) for Black non-Hispanic and White non-Hispanic Americans, respectively. The results for cardiovascular mortality were similar for these two groups. All IRRs (except with a lag of 5 years) among Hispanic Americans were below the null, indicating that mortality rates for Hispanic Americans were lower in counties with higher $PM_{2.5}$ concentrations.

In the analysis using only those counties that had a $PM_{2.5}$ monitor, we estimated IRRs that were similar to the main analysis, both in terms of point estimate and confidence interval width. We estimated that, for every 1 log ($\mu g/m^3$) increase in $PM_{2.5}$ concentration, the IRR for all-cause mortality was 1.027 (95% CI: 1.003, 1.050). The IRRs for cardiovascular and respiratory mortality were 1.083 (95% CI: 1.029, 1.141) and 1.026 (95% CI: 0.981, 1.074), respectively. Even the negative control outcome results were similar; we estimated an IRR for accidental mortality of 0.863 (95% CI: 0.753, 0.990).

DISCUSSION

In this paper, we demonstrated how Bayesian Maximum Entropy can be used to interpolate $PM_{2.5}$ concentrations from EPA monitors to the population center of US counties. We then demonstrated how the exposure estimates from this model can be used in a standard epidemiologic analysis. Specifically, we estimated the association between the interpolated $PM_{2.5}$ concentrations and rates of mortality among all residents of the contiguous US, 2000–2016. We found that rates of all-cause, non-accidental, cardiovascular, and respiratory

mortality increased as $PM_{2.5}$ concentrations increased, conditional on the sociodemographic variables included in the model. We saw this association overall and in all population subgroups, except Hispanic Americans. Associations were generally stronger for cardiovascular mortality than for all-cause mortality.

Bayesian Maximum Entropy is a powerful geostatistical method that has seen limited use in epidemiology, even though it has greater flexibility than more commonly used spatial interpolation methods (such as kriging or inverse distance weighting).[13] Our implementation here followed approaches described in the environmental modeling literature.[34,35] Bayesian Maximum Entropy is not the only way we could have obtained nationwide $PM_{2.5}$ concentrations, though. For example, we could have used the EPA's Community Multiscale Air Quality Modeling System (CMAQ) or satellite-derived data to obtain exposure estimates.[38] Nevertheless, there are key advantages of our approach. First, the underlying models and assumptions were transparent. We were able as the analysts to choose the form of the models being used (e.g., when specifying the covariance model) and many of the model parameters (e.g., the mean trend) based on substantive knowledge. We additionally were able to specifically estimate $PM_{2.5}$ concentrations at the population center of each county in the contiguous US. Other publicly available data sources generally provide air pollution concentrations on a pre-defined grid, from which we would have needed to derive county-specific estimates. Here, we were able to target the counties directly.

Second, while our soft data approach was straightforward, more advanced implementations of Bayesian Maximum Entropy can take a hybrid approach. In addition to using the concentrations recorded at EPA monitors, we could also have integrated into the model information from satellites, CMAQ, land use variables, or other sources.[36,39–42] This may have resulted in more robust exposure predictions,[43] but there are a few key limitations to incorporating additional data sources. For example, data sources like CMAQ and some satellite-derived data involve the use of modeling to incorporate information on atmospheric transport or meteorologic variables. Thus, we would have added additional modeling assumptions to those we already made. Furthermore, if we included satellite data, we would have needed to appropriately account for non-random missingness that results from frequency of measurements, cloud coverage, or surface geographic characteristics.[43] It is finally worth noting that previous studies have observed high correlations and similar cross-validation results between $PM_{2.5}$ estimates obtained using a Bayesian Maximum Entropy model that incorporated only ground-based measurements and one that incorporated both ground-based measurements and satellite data.[39,42] One paper also reported similar hazard ratio estimates for these two exposure models.[39]

The results of our analysis largely agreed with the existing literature, which has consistently found a positive association between $PM_{2.5}$ and all-cause and cause-specific mortality. However, our estimates were not as strong as those reported by Crouse et al, who estimated in a nationwide cohort of Canadian adults (randomly selected to provide detailed census data), 1999–2001, hazard ratios for a 10 $\mu g/m^3$ increase in $PM_{2.5}$ concentration of 1.15 for non-accidental mortality and 1.16 for cardiovascular mortality.[44] In comparison, we estimated IRRs of 1.038 and 1.116 for an increase from 5 $\mu g/m^3$ to 15 $\mu g/m^3$, respectively for these two categories of cause of death. Our results did not change if we subset to the

period 2000–2001 to match the Crouse study. In a nationwide Medicare cohort, followed from 2000 to 2012, Di et al estimated all-cause hazard ratios that were also stronger than those estimated here. In the single pollutant analysis, they reported a hazard ratio of 1.08 per 10 μg/m^3 increase in $PM_{2.5}$ concentration. If we restrict our analysis to those 65 years of age or older, we estimate an IRR for all-cause mortality of 1.042 (95% CI: 1.041, 1.044) per 1 log(μg/m^3) increase, which translates to an IRR of 1.046 for an increase in $PM_{2.5}$ concentration from 5 μg/m^3 to 15 μg/m^3.

Our attenuated results could be a result of our target population, data sources, or biases. For example, a lack of exchangeability between counties with different concentrations of $PM_{2.5}$ would have led to bias in the results. Indeed, ecologic analyses like the one carried out here are particularly vulnerable to residual confounding.[45, p. 522–523] If variables such as population density or the concentrations of co-pollutants were confounders (by which we mean that not controlling for these variables meant there was an open back-door path between county-level ambient $PM_{2.5}$ and mortality), we would have incurred bias by leaving them out of our models. The findings from our negative control outcome support the possibility that our main results were biased. We estimated a strong negative association between $PM_{2.5}$ and accidental mortality. If accidental mortality was a valid negative control (which is based on the assumption that $PM_{2.5}$ has no true causal effect on the causes of death we defined as accidental mortality and that the unmeasured confounders of this relationship are the same as those for non-accidental, cardiovascular, and respiratory mortality),[9,11,25] these findings suggest that our main results were biased down, although we cannot infer the magnitude of the bias without further assumptions.[23]

An additional limitation of our analysis was the potential for measurement error. Although modeling concentrations of $PM_{2.5}$ in counties without an EPA monitor allowed us to extend the population under consideration from those who live near monitors to the entire contiguous US, there will be uncertainty associated with those predicted values that will increase as distance from monitor increases. How approaches such as ours can lead to measurement error has been discussed in the literature.[46] There exist methods we could use to account for the effect of measurement error on our point estimate and its standard error; for example, parametric or nonparametric bootstrap would be suitable for our analysis due to the lack of validation data.[46–48] Two challenges to implementing a bootstrap approach were the lack of existing software to wrap Bayesian Maximum Entropy within bootstrap resamples and the computational intensity of such a process (when the exposure modeling is already computationally intensive in one iteration). Future work could focus on integrating methods for measurement error correction with Bayesian Maximum Entropy. Nonetheless, we saw similar results when we subset to those counties that had a monitor for $PM_{2.5}$, i.e. those counties where we expect exposure misclassification to be minimized in this context.

Here, we demonstrated how one can use the advanced geostatistical method Bayesian Maximum Entropy to estimate county-specific concentrations of $PM_{2.5}$, even in counties without monitors. We then demonstrated how the resulting exposure estimates could be used in a traditional epidemiologic analysis that quantified the association between county-level $PM_{2.5}$ concentrations and mortality rates in the US, without restricting our analysis to counties with monitors. This is not the only possible application of Bayesian Maximum

Entropy that could be useful for public health. This approach could also be used to predict exposure in observational cohorts that have individual geographical information, to carry out risk assessments, or to inform where new air pollution monitors should be built in low- to middle-income countries (by determining the regions that have exposure estimates with high variance).[49,50] Future work could also consider how we might appropriately use the exposure estimates obtained from Bayesian Maximum Entropy in analyses that examine how interventions on nationwide $PM_{2.5}$ might have a causal impact on US mortality rates.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments:

Data were provided by the National Center for Health Statistics (NCHS). All analyses, interpretations, or conclusions reached are credited to the authors and not to NCHS.

Source of Funding:

This work was supported by grants T32 ES007018 from the National Institute of Environmental Health Sciences; R01 HD093602, R01 AI100654, and K01 AI125087 from the National Institutes of Health; and a Gillings Innovation Laboratory award funded by the 2007 Gillings Gift to the Gillings School of Global Public Health.

REFERENCES

1. U.S. EPA. 2009 Final Report: Integrated Science Assessment for Particulate Matter. Washington, DC2009. EPA/600/R-08/139F.
2. Samet JM. The Clean Air Act and health--a clearer view from 2011. N Engl J Med. 2011;365(3):198–201. [PubMed: 21732828]
3. U.S. EPA. Integrated Science Assessment for Particulate Matter. Research Triangle Park, NC2019. EPA/600/R-19/188.
4. U.S. EPA. Progress Cleaning the Air and Improving People's Health. https://www.epa.gov/clean-air-act-overview/progress-cleaning-air-and-improving-peoples-health#pollution. Published 2018. Updated 08/14/2018. Accessed 08/27/2018.
5. Pope CA 3rd, Cropper M, Coggins J, Cohen A. Health benefits of air pollution abatement policy: Role of the shape of the concentration-response function. J Air Waste Manag Assoc. 2015;65(5):516–522. [PubMed: 25947311]
6. Marshall JD, Apte JS, Coggins JS, Goodkind AL. Blue Skies Bluer? Environ Sci Technol. 2015;49(24):13929–13936. [PubMed: 26535809]
7. Samet JM, Dominici F, Curriero FC, Coursac I, Zeger SL. Fine particulate air pollution and mortality in 20 U.S. cities, 1987–1994. N Engl J Med. 2000;343(24):1742–1749. [PubMed: 11114312]
8. Pope CA 3rd, Dockery DW, Spengler JD, Raizenne ME. Respiratory health and PM10 pollution. A daily time series analysis. Am Rev Respir Dis. 1991;144(3 Pt 1):668–674. [PubMed: 1892309]
9. Pope CA 3rd, Thun MJ, Namboodiri MM, et al. Particulate air pollution as a predictor of mortality in a prospective study of U.S. adults. Am J Respir Crit Care Med. 1995;151(3 Pt 1):669–674. [PubMed: 7881654]
10. Pope CA 3rd, Burnett RT, Thun MJ, et al. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. JAMA. 2002;287(9):1132–1141. [PubMed: 11879110]
11. Dockery DW, Pope CA 3rd, Xu X, et al. An association between air pollution and mortality in six U.S. cities. N Engl J Med. 1993;329(24):1753–1759. [PubMed: 8179653]

12. Laden F, Schwartz J, Speizer FE, Dockery DW. Reduction in fine particulate air pollution and mortality: Extended follow-up of the Harvard Six Cities study. Am J Respir Crit Care Med. 2006;173(6):667–672. [PubMed: 16424447]
13. Jerrett M, Burnett RT, Ma R, et al. Spatial analysis of air pollution and mortality in Los Angeles. Epidemiology. 2005;16(6):727–736. [PubMed: 16222161]
14. Eftim SE, Samet JM, Janes H, McDermott A, Dominici F. Fine particulate matter and mortality: a comparison of the six cities and American Cancer Society cohorts with a medicare cohort. Epidemiology. 2008;19(2):209–216. [PubMed: 18223484]
15. Zeger SL, Dominici F, McDermott A, Samet JM. Mortality in the Medicare population and chronic exposure to fine particulate air pollution in urban centers (2000–2005). Environ Health Perspect. 2008;116(12):1614–1619. [PubMed: 19079710]
16. Correia AW, Pope CA 3rd, Dockery DW, Wang Y, Ezzati M, Dominici F. Effect of air pollution control on life expectancy in the United States: an analysis of 545 U.S. counties for the period from 2000 to 2007. Epidemiology. 2013;24(1):23–31. [PubMed: 23211349]
17. Pope CA 3rd, Ezzati M, Dockery DW. Fine-particulate air pollution and life expectancy in the United States. N Engl J Med. 2009;360(4):376–386. [PubMed: 19164188]
18. Dominici F, Peng RD, Bell ML, et al. Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. JAMA. 2006;295(10):1127–1134. [PubMed: 16522832]
19. Ostro B, Lipsett M, Reynolds P, et al. Long-term exposure to constituents of fine particulate air pollution and mortality: results from the California Teachers Study. Environ Health Perspect. 2010;118(3):363–369. [PubMed: 20064787]
20. Pope CA 3rd. Particulate pollution and health: a review of the Utah valley experience. J Expo Anal Environ Epidemiol. 1996;6(1):23–34. [PubMed: 8777371]
21. Wang Y, Kloog I, Coull BA, Kosheleva A, Zanobetti A, Schwartz JD. Estimating Causal Effects of Long-Term PM2.5 Exposure on Mortality in New Jersey. Environ Health Perspect. 2016;124(8):1182–1188. [PubMed: 27082965]
22. National Center for Health Statistics. Compressed Mortality File, 1999–2016 (machine readable data file and documentation, CD-ROM Series 20, No. 2V) as compiled from data provided by the 57 vital statistics jurisdictions through the Vital Statistics Cooperative Program. In. Hyattsville, MD2016.
23. Lipsitch M, Tchetgen Tchetgen E, Cohen T. Negative controls: a tool for detecting confounding and bias in observational studies. Epidemiology. 2010;21(3):383–388. [PubMed: 20335814]
24. Bowe B, Xie Y, Yan Y, Al-Aly Z. Burden of Cause-Specific Mortality Associated With PM2.5 Air Pollution in the United States. JAMA Netw Open. 2019;2(11):e1915834. [PubMed: 31747037]
25. Krewski D, Burnett RT, Goldberg MS, et al. Overview of the reanalysis of the Harvard Six Cities Study and American Cancer Society Study of Particulate Air Pollution and Mortality. J Toxicol Environ Health A. 2003;66(16–19):1507–1551. [PubMed: 12959828]
26. U.S. Census. American Community Survey (ACS): When to Use 1-year, 3-year, or 5-year Estimates. https://www.census.gov/programs-surveys/acs/guidance/estimates.html. Published 2018. Updated September 6, 2018. Accessed.
27. SAS Institute Inc. The EXPAND Procedure. http://support.sas.com/documentation/cdl/en/etsug/68148/HTML/default/viewer.htm#etsug_expand_toc.htm. Published 2015. Accessed.
28. Serre M, Bogaert P, Christakos G. Computational investigations of Bayesian maximum entropy spatiotemporal mapping. Paper presented at: 4th Annual Conference1998.
29. Christakos G A Bayesian/maximum-entropy view to the spatial estimation problem. Mathematical Geology. 1990;22(7):763–777.
30. Christakos G, Serre M. BME Analysis of Particulate Matter Distributions in North Carolina. Atmospheric Environment. 2000;34(20):3393–3406.
31. Shannon CE. A mathematical theory of communication. Bell System Technical Journal. 1948;27:379–423.
32. Schneider TD. Information Theory Primer. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.147.2463&rep=rep1&type=pdf. Published 2007. Updated 04/14/2007. Accessed.

33. Serre ML, Christakos G. Modern geostatistics: Computational BME in the light of uncertain physical knowledge--the Equus Beds Study. Stochastic Environmental Research and Risk Assessment. 1999;13(1):1–26.
34. Akita Y, Chen JC, Serre ML. The moving-window Bayesian maximum entropy framework: estimation of PM(2.5) yearly average concentration across the contiguous United States. J Expo Sci Environ Epidemiol. 2012;22(5):496–501. [PubMed: 22739679]
35. Serre ML, Christakos G, Lee SJ. Soft Data Space/Time Mapping of Coarse Particulate Matter Annual Arithmetic Average over the U.S. In: Sanchez-Vila X, ed. geoENV IV - Geostatistics for Environmental Applications. Vol 1. Dordrecht: Kluwer Academic Publishers; 2004:115–126.
36. Reyes JM, Serre ML. An LUR/BME framework to estimate PM2.5 explained by on road mobile and stationary sources. Environ Sci Technol. 2014;48(3):1736–1744. [PubMed: 24387222]
37. Stokes ME, Davis CS, Koch GG. Using GEE to Account for Overdispersion: Univariate Outcome. In: Categorical Data Analysis Using SAS. Third Edition ed. Cary, NC: SAS Institute, Inc.; 2012:549–554.
38. U.S. EPA. CMAQ: The Community Multiscale Air Quality Modeling System. https://www.epa.gov/cmaq. Published 2021. Accessed March 29, 2021.
39. Jerrett M, Turner MC, Beckerman BS, et al. Comparing the Health Effects of Ambient Particulate Matter Estimated Using Ground-Based versus Remote Sensing Exposure Estimates. Environ Health Perspect. 2017;125(4):552–559. [PubMed: 27611476]
40. Beckerman BS, Jerrett M, Serre M, et al. A hybrid approach to estimating national scale spatiotemporal variability of PM2.5 in the contiguous United States. Environ Sci Technol. 2013;47(13):7233–7241. [PubMed: 23701364]
41. Meng J, Li C, Martin RV, van Donkelaar A, Hystad P, Brauer M. Estimated Long-Term (1981–2016) Concentrations of Ambient Fine Particulate Matter across North America from Chemical Transport Modeling, Satellite Remote Sensing, and Ground-Based Measurements. Environ Sci Technol. 2019;53(9):5071–5079. [PubMed: 30995030]
42. Cleland SE, West JJ, Jia Y, et al. Estimating wildfire smoke concentrations during the October 2017 California fires through BME space/time data fusion of observed, modeled, and satellite-derived PM2.5. Environ Sci Technol. In Press.
43. Diao M, Holloway T, Choi S, et al. Methods, availability, and applications of PM2.5 exposure estimates derived from ground measurements, satellite, and atmospheric models. J Air Waste Manag Assoc. 2019;69(12):1391–1414. [PubMed: 31526242]
44. Crouse DL, Peters PA, van Donkelaar A, et al. Risk of nonaccidental and cardiovascular mortality in relation to long-term exposure to low concentrations of fine particulate matter: a Canadian national-level cohort study. Environ Health Perspect. 2012;120(5):708–714. [PubMed: 22313724]
45. Rothman KJ, Greenland S, Lash TL. Modern Epidemiology. 3rd ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2008.
46. Szpiro AA, Sheppard L, Lumley T. Efficient measurement error correction with spatially misaligned data. Biostatistics. 2011;12(4):610–623. [PubMed: 21252080]
47. Samoli E, Butland BK. Incorporating Measurement Error from Modeled Air Pollution Exposures into Epidemiological Analyses. Curr Environ Health Rep. 2017;4(4):472–480. [PubMed: 28983855]
48. Szpiro AA, Paciorek CJ. Measurement error in two-stage analyses, with application to air pollution epidemiology. Environmetrics. 2013;24(8):501–517. [PubMed: 24764691]
49. Puangthongthub S, Wangwongwatan S, Kamens RM, Serre ML. Modeling the Space/Time Distribution of Particulate Matter in Thailand and Optimizing its Monitoring Network. Atmospheric Environment. 2007;41:7788–7805.
50. Allshouse WB, Pleil JD, Rappaport SM, Serre ML. Mass fraction spatiotemporal geostatistics and its application to map atmospheric polycyclic aromatic hydrocarbons after 9/11. Stoch Environ Res Risk Assess. 2009;23:1213–1223.

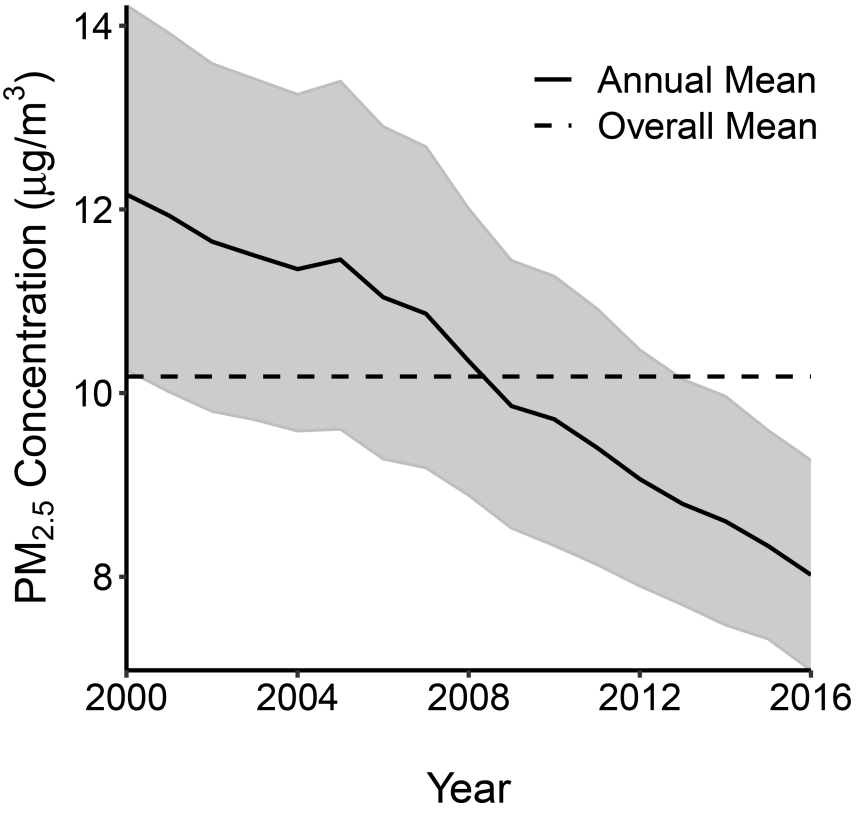

Figure 1.

Average concentrations of $PM_{2.5}$ in all counties in the contiguous US across, 2000–2016. The gray band represents the 25^{th} and 75^{th} percentiles of the year- and county-specific concentrations. $PM_{2.5}$ concentrations in $\mu g/m^3$ were derived from the Bayesian Maximum Entropy estimates, using the year- and county-specific average concentration in $\log(\mu g/m^3)$ and its variance.

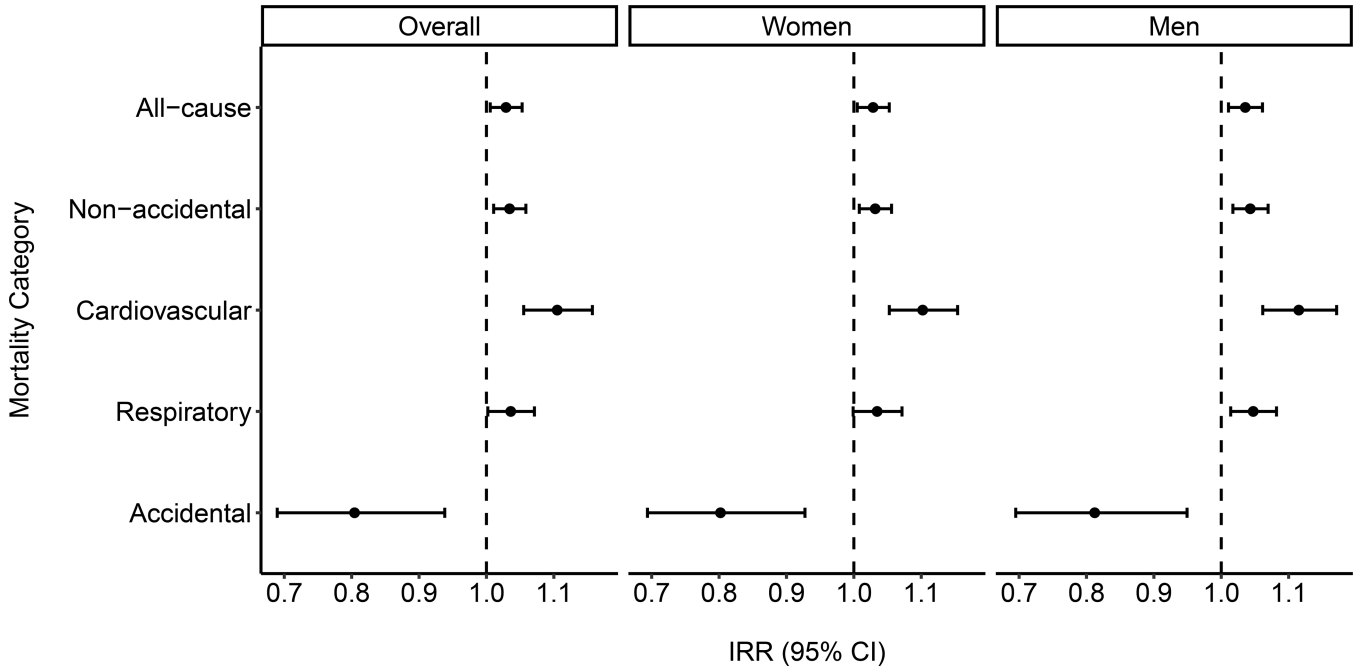

Figure 2.

Incidence rate ratios (IRR) comparing mortality rates per 1 log(μg/m^3) increase in concentrations of $PM_{2.5}$ (lag: 0 years) in all US residents and stratified by gender. CI indicates confidence interval.

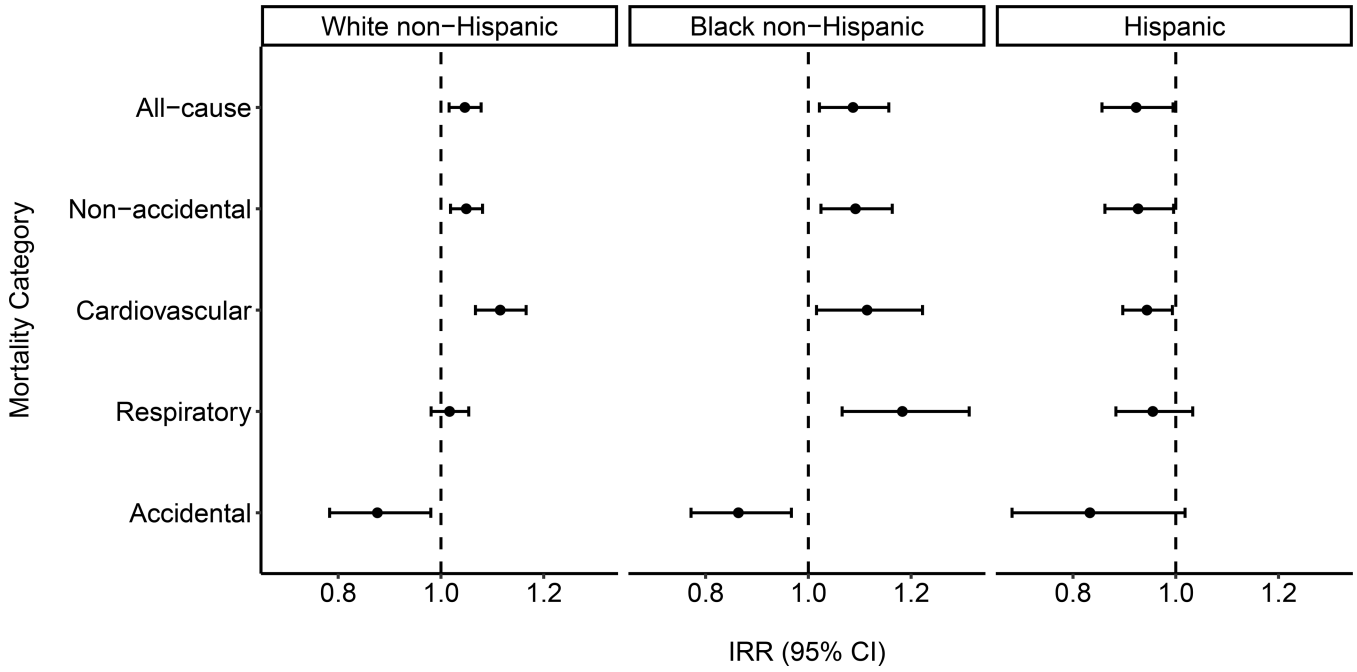

Figure 3.

Incidence rate ratios (IRR) comparing mortality rates per 1 log(μg/m^3) increase in concentrations of $PM_{2.5}$ (lag: 0 years), stratified by race-ethnicity. CI indicates confidence interval.

Table 1.

Demographic and socioeconomic characteristics of the resident population of the contiguous US, 2000–2016

Characteristic	2000–2016	2000	2016
Women (%)	50.9	51.0	50.8
Race (%)			
Black	13.6	13.1	14.1
White	80.1	4.9	78.4
Other	6.3	82.0	7.5
Hispanic ethnicity (%)	15.6	12.6	17.8
Median annual household income ($)[a]	34,883.9	34,804.0	33,726.5
Urbanicity (%)			
Mostly urban	86.2	84.8	87.6
Mostly rural	12.2	13.2	10.8
Rural	1.7	2.1	1.7
Population graduated high school (%)[b]	83.2	77.6	85.9

[a] National median of county-specific median incomes, in 2000 inflation-adjusted dollars

[b] National average of county-specific percents

Table 2.

Crude and age-adjusted[a] mortality rates (per 100,000 person–years) for all counties in the contiguous US, 2000–2016

Year(s)	All-cause		Non-accidental		Cardiovascular		Respiratory	
	Crude	Adjusted	Crude	Adjusted	Crude	Adjusted	Crude	Adjusted
2000–2016	841.5	779.9	826.9	765.5	280.0	257.3	81.4	75.4
2000	865.5	865.5	848.6	848.6	335.7	335.7	83.1	83.1
2002	862.4	853.6	845.2	836.6	322.5	319.2	82.6	81.9
2004	833.2	813.6	816.5	797.0	297.8	290.5	78.7	77.1
2006	829.0	793.4	812.4	777.0	280.2	266.8	77.2	74.2
2008	829.7	777.5	815.4	763.3	268.7	249.8	82.0	77.2
2010	817.0	750.4	804.5	738.1	257.0	233.6	78.3	72.3
2012	828.3	737.0	815.7	724.6	253.7	222.6	79.9	71.1
2014	844.3	730.5	832.1	718.7	256.8	218.7	82.7	71.2
2016	871.8	736.1	858.2	722.9	264.2	218.7	84.1	70.1

[a]Standardized to the age distribution of the 2000 resident population of the contiguous US

Table 3.

Incidence rate ratios comparing mortality rates per 1 log(μg/m^3) increase in lagged concentrations of $PM_{2.5}$ in all US residents and stratified by gender.

Exposure Lag (Years)	Cause of Death	All		Women		Men	
		IRR	95% CI	IRR	95% CI	IRR	95% CI
0	All causes	1.029	1.006, 1.053	1.029	1.005, 1.053	1.036	1.011, 1.061
	Non-accidental	1.034	1.011, 1.059	1.032	1.008, 1.056	1.043	1.017, 1.070
	Cardiovascular	1.105	1.055, 1.157	1.102	1.053, 1.154	1.115	1.062, 1.171
	Respiratory	1.036	1.002, 1.071	1.035	0.999, 1.071	1.047	1.014, 1.082
	Accidental	0.804	0.690, 0.938	0.802	0.694, 0.927	0.812	0.695, 0.949
1	All causes	1.024	1.001, 1.046	1.022	1.000, 1.044	1.031	1.007, 1.056
	Non-accidental	1.029	1.006, 1.051	1.025	1.003, 1.047	1.038	1.014, 1.063
	Cardiovascular	1.098	1.050, 1.147	1.093	1.047, 1.142	1.109	1.057, 1.163
	Respiratory	1.027	0.993, 1.064	1.024	0.988, 1.061	1.040	1.006, 1.076
	Accidental	0.806	0.687, 0.947	0.804	0.692, 0.933	0.815	0.692, 0.959
3	All causes	1.013	0.991, 1.036	1.011	0.989, 1.035	1.020	0.998, 1.043
	Non-accidental	1.018	0.996, 1.040	1.014	0.992, 1.036	1.026	1.004, 1.049
	Cardiovascular	1.085	1.040, 1.131	1.081	1.038, 1.126	1.094	1.043, 1.146
	Respiratory	1.018	0.977, 1.062	1.019	0.976, 1.064	1.025	0.985, 1.066
	Accidental	0.811	0.690, 0.953	0.806	0.690, 0.941	0.819	0.697, 0.964
5	All causes	1.009	0.982, 1.037	1.006	0.978, 1.035	1.017	0.991, 1.043
	Non-accidental	1.013	0.988, 1.039	1.009	0.982, 1.036	1.022	0.998, 1.047
	Cardiovascular	1.076	1.033, 1.122	1.073	1.030, 1.118	1.084	1.037, 1.134
	Respiratory	1.019	0.969, 1.073	1.021	0.968, 1.077	1.023	0.975, 1.074
	Accidental	0.821	0.696, 0.969	0.818	0.696, 0.961	0.829	0.703, 0.978

Abbreviations: $PM_{2.5}$, particulate matter with an aerodynamic diameter <2.5 μm; IRR, incidence rate ratios; CI, confidence interval

Table 4.

Incidence rate ratios comparing mortality rates per 1 log(μg/m^3) increase in lagged concentrations of $PM_{2.5}$ stratified by race-ethnicity.

Exposure Lag (Years)	Cause of Death	White non-Hispanic		Black non-Hispanic		Hispanic	
		IRR	95% CI	IRR	95% CI	IRR	95% CI
0	All causes	1.047	1.016, 1.078	1.087	1.021, 1.157	0.923	0.856, 0.995
	Non-accidental	1.049	1.019, 1.081	1.092	1.025, 1.163	0.927	0.862, 0.996
	Cardiovascular	1.115	1.166, 1.067	1.114	1.016, 1.222	0.944	0.897, 0.993
	Respiratory	1.017	0.981, 1.054	1.183	1.066, 1.313	0.955	0.883, 1.033
	Accidental	0.876	0.783, 0.980	0.864	0.772, 0.967	0.833	0.681, 1.018
1	All causes	1.040	1.010, 1.071	1.072	1.009, 1.139	0.927	0.862, 0.998
	Non-accidental	1.043	1.013, 1.074	1.077	1.012, 1.146	0.931	0.868, 0.999
	Cardiovascular	1.108	1.061, 1.156	1.099	1.004, 1.203	0.960	0.914, 1.008
	Respiratory	1.007	0.970, 1.046	1.160	1.046, 1.286	0.972	0.899, 1.050
	Accidental	0.873	0.775, 0.982	0.857	0.757, 0.971	0.833	0.676, 1.026
3	All causes	1.030	1.000, 1.061	1.039	0.981, 1.101	0.934	0.869, 1.004
	Non-accidental	1.032	1.003, 1.063	1.043	0.984, 1.106	0.938	0.875, 1.005
	Cardiovascular	1.095	1.051, 1.141	1.071	0.978, 1.172	0.985	0.930, 1.042
	Respiratory	0.998	0.953, 1.046	1.116	1.008, 1.236	0.987	0.921, 1.057
	Accidental	0.870	0.770, 0.983	0.845	0.743, 0.962	0.832	0.679, 1.021
5	All causes	1.025	0.993, 1.059	1.012	0.954, 1.074	0.951	0.886, 1.020
	Non-accidental	1.028	0.996, 1.060	1.016	0.957, 1.078	0.954	0.891, 1.022
	Cardiovascular	1.085	1.040, 1.132	1.046	0.956, 1.144	1.029	0.962, 1.101
	Respiratory	0.999	0.944, 1.058	1.073	0.969, 1.189	1.016	0.958, 1.077
	Accidental	0.871	0.765, 0.991	0.840	0.722, 0.976	0.852	0.706, 1.028

Abbreviations: $PM_{2.5}$, particulate matter with an aerodynamic diameter <2.5 μm; IRR, incidence rate ratios; CI, confidence interval