Science Advances
AAAS

Supplementary Materials for

Air pollution and COVID-19 mortality in the United States: Strengths and limitations of an ecological regression analysis

X. Wu,[1*] R. C. Nethery,[1*] M. B. Sabath,[1] D. Braun,[1,2] F. Dominici[1†]

*These authors contributed equally to this work.
†Corresponding author. Email: fdominic@hsph.harvard.edu

This PDF file includes:

Materials and Methods
Section S1. Data
Section S2. Statistical Methods
Section S3. Analysis Results
Section S4. Analysis Discussion
Section S5. Code
Fig. S1. COVID-19 mortality rate ratios (MRR) per 1 μg/m^3 increase in $PM_{2.5}$ and 95% CI for the analyses using four relevant sets of exposure data.
Fig. S2. COVID-19 mortality rate ratios (MRR) per empirical quintile increase in $PM_{2.5}$ and 95% CI for the analyses using four relevant sets of exposure data.
Fig. S3. COVID-19 mortality rate ratios (MRR) per 1 μg/m^3 increase in $PM_{2.5}$ and 95% CI using daily cumulative COVID-19 death counts from April 18, 2020 to June 18, 2020.
Table S1. Publicly available data sources used in the analyses
Table S2. Characteristics of the study cohort up to and including June 18, 2020
Table S3. Main, secondary, and sensitivity analysis results for exposure data P-1
Table S4. Main, secondary, and sensitivity analysis results for exposure data P-2
Table S5. Main, secondary, and sensitivity analysis results for exposure data P-3
Table S6. Main, secondary, and sensitivity analysis results for exposure data P-4
References (28-59)

Materials and Methods

S.1 Data

COVID-19 deaths

We obtained COVID-19 death counts for each county in the United States from Johns Hopkins University, Center for Systems Science and Engineering Coronavirus Resource Center (28). This source provides the most comprehensive county-level COVID-19 data to date reported by the US Centers for Disease Control and Prevention (CDC) and state health departments, including the number of new and cumulative deaths and confirmed cases reported in each county across the United States, updated daily. We collected the cumulative number of deaths for each county up to and including June 18, 2020. County-level COVID-19 mortality rates were defined for our analyses as the ratio of COVID-19 deaths to county-level population size. While individual-level data would have allowed a more rigorous statistical analyses, individual-level data on COVID-19 death is currently not available. As of May 11, 2020, the CDC reports that COVID-19 testing is being conducted at 97 public health laboratories across the US and territories. The CDC website states that COVID-19 deaths are identified using the International Statistical Classification of Diseases and Related Health Problems (ICD) codes for cause of death recorded on death certificates. The ICD-10 code indicating a COVID-19 related death is U07.1. The CDC notes that deaths reported with this code "can include cases with or without laboratory confirmation." They also note that the data may be affected by delays in reporting and by differential reporting practices across states. Also, there are numbers of excess deaths associated with COVID-19 especially during the early stage of the pandemic.

Exposure to air pollution

We calculated county-level long-term exposure to $PM_{2.5}$ (averaged from 2000 to 2016) from an established exposure prediction model (9). The $PM_{2.5}$ exposure levels were estimated monthly at $0.01^\circ \times 0.01^\circ$ grid resolution across the entire continental United States by fusing $PM_{2.5}$ measures from three different sources: ground-based monitors, GEOS-Chem chemical transport models (CTM), and satellite observations. In short, CTM and satellite data are combined to estimate a high-resolution $PM_{2.5}$ surface across the whole United States, then this surface is bias-corrected for ground-monitor $PM_{2.5}$ observations using a geographically-weighted regression. These estimates have been extensively cross-validated and the cross-validated R^2 for these models in the United States was reported to be 0.61, although the accuracy varies across regions (9). We aggregated these levels spatially by averaging the values for all grid points within a zip code and then averaging across zip codes within a county. We obtained temporally averaged $PM_{2.5}$ values (2000–2016) at the county level by averaging estimated $PM_{2.5}$ values within a given county. We computed the average 2016 $PM_{2.5}$ exposure analogously for each county to use in sensitivity analyses.

To assess the sensitivity of our results to the specific $PM_{2.5}$ prediction model used to generate exposure estimates, we also collect the estimated daily $PM_{2.5}$ modeled exposure at a high spatio-temporal resolution of 1 km $\times$ 1 km grid network across the whole United States using another well-validated ensemble-based prediction model (29). This model used ensemble learning approaches to combine three machine learning models; a random forest regression, a gradient boosting machine, and an artificial neural network. These machine learning algorithms used more than 100 predictor variables from satellite data, land-use information, weather variables, and output from chemical transport model simulations. We use the same area-aggregation approach to aggregate the gridded data to the county-level and then were averaged across the years 2000-2016.

Potential confounders

To adjust for confounding bias in the nationwide observational study, we use county-level variables from numerous public sources. In the main analysis, we considered the following 19 county-level variables and one state-level variable as potential confounders (see also Table S2): days since first COVID-19 case reported (a proxy for epidemic stage), population density, percent of population $\geq$65 years of age, percent of the population 45-64 years of age, percent of the population 15-44 years of age, percent of the population living in poverty, median household income, percent of Black residents, percent of Hispanic residents, percent of the adult population with less than high school education, median house value, percent of owner-occupied housing, percent of the population with obesity, percent of current smokers, number of hospital beds per unit population, and average daily temperature and relative humidity for summer (June to September) and winter (December to February) for each county, and days since issuance of stay-at-home order for each state. Note that publicly available daily COVID-19 case counts at the county-level were only available starting March 22, 2020, so that the measure of days since first COVID-19 case reported was truncated by this date.

Multiple socioeconomic and demographic variables were collected from the 2000 and 2010 Census, the 2005–2016 American Community Surveys and the 2009–2016 CDC Compressed Mortality File. Specifically, we collect and calculate the following 11 county-level socioeconomic and demographic variables: proportion of residents older than 65 years of age, proportion of residents 45-64 years of age, proportion of residents 15-44 years of age, proportion of Hispanic residents, proportion of Black residents, median household income, median home value, proportion of residents living in poverty, proportion of the adult residents with less than high school education, population density, and proportion of residents that own their house. The American Community Survey data were extracted for each zip code and then averaging across zip codes within a county. We also collect two county-level health risk factors: proportion of residents with obesity and proportion of residents that are current smokers from the Robert Wood Johnson Foundation's 2020 County Health Rankings.

Certain features of the counties' COVID-19 outbreaks and response and the accessibility of health care may also confound the relationship between $PM_{2.5}$ and COVID-19 mortality. One particularly important feature is the county's point on the epidemic curve at the time of analysis. Although this feature is difficult to accurately measure, we approximate it using time since first reported COVID-19 case. This information is also extracted from the JHU-CSSE database (the same source used for the COVID-19 death counts). States also issued "stay-at-home/shelter-in-place" orders in response to the outbreak at different times, which likely affected infection rates and could also be associated with $PM_{2.5}$. Thus, we also adjust our models for state-level time since implementation of stay-at-home/shelter-in-place order, obtained from COVID-19 US state policy database. During the course of COVID-19 outbreak, the availability of adequate hospital resources and of testing resources likely influence COVID-19 outcomes and these may also be more widely available in urban areas where $PM_{2.5}$ is also higher. We collect county-level information on number of hospital beds available in 2019 from Homeland Infrastructure Foundation-Level Data (HIFLD) and state-level information on number of COVID-19 tests performed up to June 18, 2020 from the COVID tracking project.

Meteorological variables are commonly adjusted for when studying the health impacts of air pollution. We obtain meteorological variables on maximum daily temperature and relative humidity data on 4 km $\times$ 4 km gridded rasters from gridMET via Google Earth Engine. We average daily temperature and relative humidity for the summer (June to September) and winter (December to February) period respectively across the period 2000-2016 and average across grid rasters in each county. We adjust for all four of these weather variables in our main models. The data used for this study are publicly available and sources are listed in Table S1.

S.2 Statistical Methods

We fit a negative binomial mixed model (30-32) using COVID-19 deaths as the outcome and $PM_{2.5}$ as the exposure of interest to estimate the association between COVID-19 mortality rate and long-term $PM_{2.5}$ exposure, adjusted by covariates. The model included a population size offset and was adjusted for all the potential confounders listed above. We also included a random intercept by state to account for potential correlation in counties within the same state, due to similar socio-cultural, behavioral, and healthcare system features and similar COVID-19 response and testing policies. We report mortality rate ratios (MRR), i.e., exponentiated parameter estimates from the negative binomial model, and 95% confidence interval (CI). The MRR for $PM_{2.5}$ can be interpreted as the relative increase in the COVID-19 mortality rate associated with a 1 $\mu g/m^3$ increase in long-term average $PM_{2.5}$ exposure. We carried out all analyses in R statistical software and performed model fitting using the lme4 package (33) For our main and secondary analyses, we fit Negative Binomial regression models with a state-specific random intercept. All potential confounders are centered and scaled prior to analysis. Letting $E[\cdot]$ denote an expected value, the main model takes the form

log[E(COVID-19 deaths)] = β_0 + β_1 $PM_{2.5}$ + β_2 population density + β_3 percent of the population $\geq$ 65 + β_4 percent of the population 45-64 + β_5 percent of the population 15-44 + β_6 percent living in poverty + β_7 median household income + β_8 percent black + β_9 percent hispanic + β_{10} percent of adults with less than a high school education + β_{11} median house value + β_{12} percent of owner-occupied housing + β_{13} percent obese + β_{14} percent smokers + β_{15} days since first case + β_{16} days since stay at home order + β_{17} number of hospital beds + β_{18} average summer temperature + β_{19} average summer relative humidity + β_{20} average winter temperature + β_{21} average winter relative humidity + offset[log(population size)] + random intercept(State)

We report the mortality rate ratios (MRR) and 95% CIs for $PM_{2.5}$, corresponding to the

exponentiated parameter estimate ($e^{\hat{\beta}_1}$). The MRR can be interpreted as the multiplicative increase in the COVID-19 death rate associated with a 1 μg/m^3 increase in long-term average $PM_{2.5}$ exposure. The statistical code in R is publicly available.

Model Assumption Diagnostics

Poisson regression models are a common choice for modeling count data, but the Poisson distribution is restrictive in that it assumes that the mean is equal to the variance. In our setting, because most counties have experienced few or no COVID-19 deaths thus far, the mean of our outcome data is small ($\mu = 37.79$); however the variance is large due to the large death counts in several outbreak epicenters ($\sigma^2 = 186934.90$). Among the counties with non-zero deaths, the mean of our outcome data is still relative small ($\mu = 63.28$); however the variance is large ($\sigma^2 = 311431.70$). The dispersion parameter for the quasi-Poisson family is estimated to be 29.01, which indicates substantial over-dispersion. Thus, the Poisson distributional assumption is likely to be inappropriate. The negative binomial distribution provides more flexibility by introducing an additional parameter that allows the count outcome variable with variance larger than mean. This flexibility also better accounts for the large number of zeros in our outcome, without requiring the use of zero-inflated models, which are more complex and less interpretable.

To assess the model fit of the standard Negative Binomial regression model comparing to a zero-inflated Negative Binomial regression model, we conduct an Vuong closeness test for the goodness-of-fit (34). We found no statistically significant improvement of model fit for a zero-inflated Negative Binomial regression model (P-Value = 0.84). The AIC of the standard Negative Binomial regression model is even slightly lower (AIC= 15012.8 vs. 15014.9), which suggests a better model fit.

Quantifying Unmeasured Confounding Bias

Because this study is observational and the contributing factors to COVID-19 spread and severity remain largely unknown at this early stage of the pandemic, unmeasured confounding is a concern in our analyses. The E-value is a commonly used metric to evaluate the potential impact of unmeasured confounding on results from an observational study (35). For a pre-specified exposure variable of interest (long-term exposure to $PM_{2.5}$), the E-value quantifies the minimum strength of association that an unmeasured confounder must have, with both the outcome (COVID-19 mortality rate) and exposure (long-term exposure to $PM_{2.5}$) conditional on all of the potential confounders included in the regression model, to explain away the estimated exposure-outcome relationship. We report the E-value for the MRR estimate for $PM_{2.5}$ under the main model with 20 potential confounders. We calculated the E-values for our reported MRRs per 1 $\mu g/m^3$ increase of long-term exposure to $PM_{2.5}$. The calculation of E-values can be implemented through the E-value calculator by Mathur et al. (36), available at `https://www.evalue-calculator.com/`.

Secondary Analyses

In addition to the main analysis, we conducted seven secondary analyses to assess the robustness of our results to the confounder set used, outliers, and the model specifications.

First, because the New York metropolitan area has experienced the most severe COVID-19 outbreak in the United States to date, we anticipated that it would strongly influence our analysis. As a result, we repeated the analysis excluding the counties comprising the New York metropolitan area, as defined by the Census Bureau.

Second, although in our main analysis we adjusted for days since first COVID-19 case reported to capture the size of an outbreak in a given county, this measure is imprecise. To further investigate the potential for residual confounding bias (i.e., if counties with high $PM_{2.5}$ exposure also tend to have large outbreaks relative to the population size, then their death rates per unit population could appear differentially elevated, inducing a spurious

correlation with $PM_{2.5}$), we also conducted analyses excluding counties with fewer than 10 confirmed COVID-19 cases.

Third, we omitted an anticipated strong confounder, days since first COVID-19 case reported, from the model. Fourth, we additionally adjusted our models for the number of tests performed at the state level to evaluate how state-level differences in testing policies might affect our results. Fifth, we additionally adjusted our models for county-level estimated percentage of people with COVID-19 symptoms to evaluate how the size of the outbreak in each county might affect our results. Sixth, we additionally adjusted our models for the longitude and latitude of the centroid of each county to evaluate how potential spatial residual confounding might affect our results. Seventh, we introduced $PM_{2.5}$ into our models as a categorical variable, categorized at the empirical quintiles, to assess the sensitivity of our results to the assumption of a linear effect of $PM_{2.5}$ on COVID-19 mortality rates.

Sensitivity Analyses

We conducted more than 80 sensitivity analyses to assess the robustness of our results to data and modeling choices. First, we repeated all the analyses using alternative methods to estimate exposure to $PM_{2.5}$ (29). Second, we fit the models, modifying the adjustment for confounders, such as using a log transformation or categorized versions of some of the covariates. Third, because our study relies on observational data, our results could be sensitive to modeling choices (e.g., distributional assumptions or assumptions of linearity). We evaluated the sensitivity to such choices by considering alternative model specifications and by fitting models stratified by county urban-rural status. Fourth, we repeat our analysis daily from April 18, 2020 to June 18, 2020 to evaluate the sensitivity to the temporal changes during the COVID-19 pandemic.

S.3 Analysis Results

Table S1 summarizes our data sources and their provenance, including links where the raw data can be extracted directly. Table S2 describes the data used in our analyses.

In Figure S1-S2, we report the MRR and 95% CI for $PM_{2.5}$ from all secondary analyses. In these analyses, we separately (a) omitted New York metropolitan area; (b) excluded counties with fewer than 10 confirmed COVID-19 cases; (c) omitted time since first reported COVID-19 case from the model; (d) additionally adjusted the model for number of tests performed; (e) additionally adjusted the model for estimated percentage of people with COVID-19 symptoms; (f) additionally adjusted the model for the longitude and latitude; and (g) treated $PM_{2.5}$ as a categorical variable. The results of these analyses were consistent with the main analysis. For the analysis of the $PM_{2.5}$ categorized into quintiles, the MRR for the k-th quintile can be interpreted as the increase in COVID-19 mortality rate associated with a change from the first quintile to the k-th quintile in long-term $PM_{2.5}$ exposure. The MRR estimates from this model monotonically increased as $PM_{2.5}$ increased, supporting the assumption of a linear relationship between $PM_{2.5}$ and COVID-19 mortality rates.

For our main analysis, we found that the E-value for the estimated MRR for $PM_{2.5}$ was 1.46. That is, for an unmeasured confounder U to fully account for the estimated effects of the long-term $PM_{2.5}$ exposure (E) on the COVID-19 mortality (Y), it would have to be associated with both long-term $PM_{2.5}$ exposure and with mortality by a risk ratio of at least 1.46-fold each, through pathways independent of all covariates already included in the model. In other words, if we were to adjust for this U in our model, the estimated MRR for $PM_{2.5}$ would be reduced to 1 (the null value). A 1.46 risk ratio means that U would need to meet the following two criteria: 1) a 1-unit increase in U would need to lead to a 46% increase in the risk of mortality (Y); *and* 2) a 1-unit increase in U would need to be associated with a 46% increase in $PM_{2.5}$ exposure levels. To get a sense of the magnitude of the required confounding effect, we also computed the E-value for some of our key measured confounders for comparison. The E-values for days since first COVID-19 case reported (1.15),

the weather variables (1.08), number of hospital beds (1.00), the behavioral risk factors (1.01) and the longitude/latitude (1.10) were significantly smaller than the reported E-values for the required unmeasured confounder. This suggests that any unmeasured confounder would need to have a confounding effect substantially larger than any of our observed confounders in order to explain away the relationship between $PM_{2.5}$ and COVID-19 mortality rate.

Alternative $PM_{2.5}$ Estimates

To evaluate the sensitivity of our results to the approach used to calculate long-term pollution exposure measure, we repeat our analyses using four relevant sets of exposure data. Using the modeled exposure estimates of van Donkelaar et al. (9), we test the 17-year average concentrations (2000-2016), i.e., the primary analysis results, and the one-year average concentrations using the most recent available year (2016), and we refer to the analyses using these exposures as P-1 and P-2, respectively. Using the modeled exposure estimates of Di et al. (29), we test the 17-year average concentrations (2000-2016) and the one-year average concentrations using the most recent available year (2016), refered to as P-3 and P-4 respectively. In each analysis, we adjust for the set of potential confounders described in Section **S.**1. The finding that long-term exposure to $PM_{2.5}$ is positively associated with increased COVID-19 mortality holds regardless of which pollution data are used. When adjusted for the full confounder set, analyses using 17-year average concentrations (2000-2016) (P-1 and P-3) give consistent point estimates for $PM_{2.5}$ and attain statistical significance using different pollution data sources. We note the analysis results using P-2 and P-4 are less agreed with each other, i.e., P-2 is notably lower, yet P-4 is notably higher, though both still attain statistical significance. Because the focus of our study is to assess the cumulative chronic effect of long-term exposure to $PM_{2.5}$, we use 17-year mean exposure data in our main report.

Differing Confounder Sets

For each of these pollution data sources, we evaluate the model sensitivity to the set of confounders adjusted for by individually omitting each of the following from the confounder set: 1) days since first reported COVID-19 case; 2) number of hospital beds in the county; 3) behavioral risk factors, i.e., population obesity rate and percent of population who are current smokers; and 4) meteorological (weather) variables, i.e., summer and winter temperature and relative humidity. We also conduct analyses adjusting for the following additional variables (separately): 1) the total number of COVID-19 tests performed up to June 18, 2020 in each state; 2) the estimated percentage of people with COVID-19 symptoms in each county; and 3) the longitude and latitude of each county. Effect estimates are presented as mortality rate ratio (MRR) per 1 μg/m^3 increase in annual $PM_{2.5}$.

We find consistent positive associations between long-term exposure to $PM_{2.5}$ and increased mortality for COVID-19 in these analyses, with MRR between $1.09 - 1.13$ across P-1 models that adjust for different potential confounders (similar results for P-2, P-3, and P-4). The removal of days since first reported COVID-19 case from the confounder set consistently elevates the $PM_{2.5}$ point estimates. This suggests that days since first reported COVID-19 case is a strong confounder, as it captures the different stages on the epidemic curve of each county.

Assessing the Impact of Outbreak Size

To evaluate the possible impact of confounding bias due to epidemic outbreak sizes, which are not accurately captured by current data, we conduct analyses 1) excluding counties in New York metropolitan area where the major outbreak is happening 2) excluding counties with less than 10 confirmed COVID-19 cases. In the analysis that excludes counties in New York metropolitan area, we still find a statistically significant association between long-term exposure to $PM_{2.5}$ and increased mortality for COVID-19 with MRR 1.11 and 95% CI $(1.05, 1.17)$ for P-1. In the analysis that excludes counties with less than 10 confirmed

COVID-19 cases, we also find a similar statistically significant association with increased mortality of COVID-19 with magnitude of MRR 1.11 and 95% CI $(1.05, 1.17)$ for P-1.

Differing Model Specifications

To evaluate the sensitivity to modeling choices (e.g., distributional assumptions or assumptions of linearity), we conduct sensitivity analyses by 1) treating $PM_{2.5}$ as a categorical variable (categorized at empirical quintiles), 2) adjusting for population density under a logarithm transformation rather than a categorical variable, 3) using a zero-inflated negative binomial mixed model, 4) using a fixed effect negative binomial model, 5) adjusting for population size as a covariate, rather than as an offset. We also conducted stratified analyses based on county urban/rural status. Counties' classifications are obtained from the National Center for Health Statistics Urban-Rural Classification Scheme for Counties (37), which assigns each US county to one of six urban-rural categories: Large central metro, Large fringe metro, Medium metro, Small metro, Micropolitan, and Non-core. Based on this, we create a two-level urban/rural variable, with Micropolitan and Non-core defined as rural, and all other types defined as urban. We then conduct the analysis separately for urban counties and rural counties.

In the analysis that treats $PM_{2.5}$ as a categorical variable, we found the magnitude of the MRRs increases dramatically and monotonically as the quintile of $PM_{2.5}$ exposures increases for P-1. Similar results are found when using P-2, P-3, and P-4. Such findings suggest that the assumption of a linear effect of $PM_{2.5}$ on COVID-19 mortality rate is reasonable and that there is no threshold for the effect of long-term exposure to $PM_{2.5}$ on COVID-19 mortality. In the analysis that adjusts for population density under a logarithm transformation, we again find a consistent statistically significant positive association with increased COVID-19 mortality. In the analysis that uses a negative binomial model accounting for zero-inflation, we find identical results as of our main analyses. In the analysis that uses a fixed effect negative binomial model, we find a very similar effect size compared to our main analyses.

In the analysis that adjusts for population size directly, rather than as an offset, we find long-term exposure to $PM_{2.5}$ is still significantly positively associated with the number of COVID-19 deaths, although here the MRR refers to the increase in the mortality count ratio of COVID-19 per unit increase of $PM_{2.5}$, rather than the increase in the mortality rate ratio.

The results of the urban/rural stratified analyses suggest an interesting pattern– the positive association between $PM_{2.5}$ and COVID-19 mortality appears to be approximately the same in rural counties [MRR 1.07, 95% CI $(0.98, 1.16)$] and in urban counties [MRR 1.09, 95% CI $(1.02, 1.16)$] for P-1. This confirms that our results are not dominated only by large urban areas where the most severe outbreaks have been reported (and often tend to be highly polluted). While the results from the rural stratified analyses are not statistically significant, likely due to smaller sample sizes and lower power, the magnitude of the effect sizes is large.

Differing Study Time

To evaluate the sensitivity to the temporal changes in many aspects during the COVID-19 pandemic, we repeat our analysis using daily cumulative COVID-19 death counts from April 18, 2020 to June 18, 2020. We found a very consistent association between long-term exposure to $PM_{2.5}$ and COVID-19 mortality throughout the time period (See Figure S3). In particular, these findings suggest that a) the different epidemic stages vary spatially and temporally, b) the sparse super-spreading events take place randomly, and c) the excess deaths in the early stage of the pandemic may not be able to cause strong confounding biases that can explain away this robust association between historical long-term exposure to $PM_{2.5}$ and COVID-19 mortality. Note there is a slightly upward trend of the MRR estimates starting since May 20, 2020. We hypothesis this may be due to a) the early excess deaths occurs more often in areas with both higher COVID-19 mortality and higher air pollution, e.g., New York metropolitan area (23), and thus leads to downward bias in the MRR estimates in earlier days; b) the lift of social distancing measures (e.g., stay-at-home-

order) in certain areas may confound the results in later days. Future investigation is needed to provide reasons behind this upward trend, yet here we conclude that the association is consistent and remains statistically significant throughout the whole time period from April 18, 2020 to June 18, 2020.

S.4 Analysis Discussion

This is the first nationwide study in the United States to estimate the relationship between long-term exposure to $PM_{2.5}$ and COVID-19 death rates. We found statistically significant evidence that an increase of 1 $\mu g/m^3$ in long-term $PM_{2.5}$ exposure is associated with an 11% (95% CI, 6 to 17%) increase in the county's COVID-19 mortality rate. Our results were adjusted for a large set of socioeconomic, demographic, weather, behavioral, epidemic stage, social isolation measures, and healthcare-related confounders and demonstrated robustness across a wide range of sensitivity analyses.

In our previous study of 60 million Americans older than 65 years of age, we found that a 1 $\mu g/m^3$ in long-term $PM_{2.5}$ exposure is associated with a 0.73% increase in the rate of all-cause mortality (25). Therefore, the same small increase in long-term exposure to $PM_{2.5}$ led to an increase in the COVID-19 death rate of a magnitude 15 times that estimated for all-cause mortality. Although the epidemiology of COVID-19 is evolving, there is a large overlap between causes of death in COVID-19 patients and the conditions caused and/or exacerbated by long-term exposure to fine particulate matter ($PM_{2.5}$). The Global Burden of Disease Study identified air pollution as a risk factor for total and cardiovascular disease mortality, and it is believed to have contributed to nearly 5 million premature deaths worldwide in 2017 alone (38). The association between $PM_{2.5}$ and health, including both infectious and chronic respiratory diseases, cardiovascular diseases, neurocognitive disease, and pregnancy outcomes in the United States and worldwide is well established (16, 25, 39-43). A recent study by our group also documented a statistically significant association between long-term

exposures to PM2.5 and ozone and risk of ARDS among older adults in the United States (44).

Our results are consistent with previous findings that air pollution exposure increases severe outcomes during infectious disease outbreaks. Ciencewicki and Jaspers (16) provide a review of the epidemiologic and experimental literature linking air pollution to infectious disease. During the 2003 outbreak of Severe Acute Respiratory Syndrome (SARS), a type of coronavirus closely related to COVID-19, Cui et al. (45) reported that locations in China with a moderate or high long-term air pollution index (API) had SARS case fatality rates 126% and 71% higher, respectively, than locations with low API under an ecological study. Long-term particulate matter exposure has been associated with hospitalizations for pneumonia in the well-controlled quasi-experimental conditions provided by the closing of the Utah Valley Steel Mill (46), and a link between long-term $PM_{2.5}$ exposure and pneumonia and influenza deaths were reported in a well-validated cohort study (4). Several studies have reported associations between short-term $PM_{2.5}$ exposure and poor infectious disease outcomes (47, 48), including higher hospitalization rates or increased medical encounters for influenza, pneumonia, and acute lower respiratory infections. In these studies, and in the literature on the association between air pollution and chronic disease outcomes, relationships with long-term pollution exposure tend to be stronger than relationships with short-term exposure (25, 49, 50), and the large effect estimate in our study is consistent with this trend. Relationships have also been detected between pollution exposures and severe outcomes in the context of past pandemics. Studies found particulate matter exposure to be associated with the mortality during the H1N1 influenza pandemic in 2009 (51, 52). Recent studies have even used historic data to show a relationship between air pollution from coal burning and mortality in the 1918 Spanish influenza pandemic (53).

With regards to the potential biological mechanisms that may explain the relationship between air pollution and viral outcomes, $PM_{2.5}$ exposure is known to be associated with many of the cardiovascular and respiratory comorbidities that dramatically increase the risk

of death in COVID-19 patients. We hypothesize that the effects captured here are largely mediated by these comorbidities and pre-existing PM-related inflammation and cellular damage (50, 54), as suggested by a recent commentary (7). Experimental studies (16, 55-57) also suggest that exposure to pollution can suppress early immune responses to the infection, leading to later increases in inflammation and worse prognosis, which may also explain our findings. Some studies (58, 59) have suggested that air pollution can also proliferate the transmission of infectious disease. If COVID-19 spread is indeed impacted by air pollution levels, which is not yet known, some of the effects detected in our study could be mediated by this factor as well.

Our analysis provides a timely characterization of the county-level relationship between historical exposure to air pollution and COVID-19 deaths in the United States. The analysis relies on up-to-date county-level COVID-19 data and well-validated air pollution exposure measures. The inability to accurately quantify the number of COVID-19 cases due to limited testing capacity prevent us from calculating accurate case fatality rates. We instead used total population size as the denominator for our mortality rates, and we additionally adjusted our models for numerous anticipated proxies of outbreak size, including time since first reported COVID-19 case, time since stay-at-home order was issued, and population density.

Throughout the analysis, we focus on the strengths and limitations of the ecological regression study design and highlight challenges due to data quality, confounding bias, exposure and outcome misclassification. Although results were robust to sensitivity analyses, the inherent limitations of the ecological study design prevent us from making statements about individual-level associations and about causality.

S.5 Code

We provide code for all analyses reported in the paper. The code can be found on `https://github.com/wxwx1993/PM_COVID`.

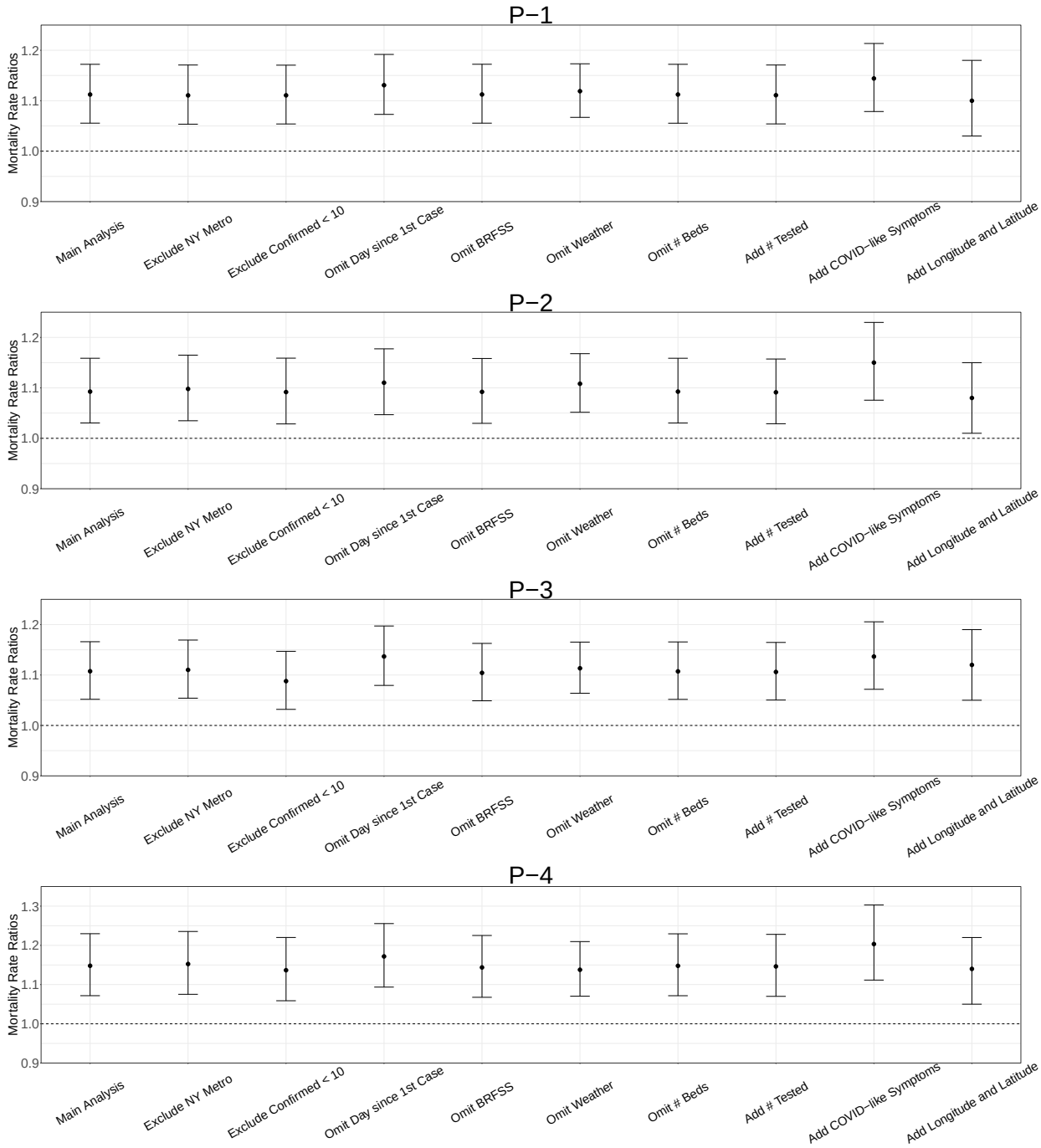

Figure S1: COVID-19 mortality rate ratios (MRR) per 1 μg/m^3 increase in $PM_{2.5}$ and 95% CI. The main analyses were adjusted for 20 socioeconomic, demographic, behavioral, meteorological, and healthcare confounders. We fit models excluding counties from NY metropolitan area, and excluding counties with < 10 confirmed cases. We conduct analyses omitting the following variables from the adjustment set: days since first COVID-19 case (day since 1st case), smoking rate and obesity rate (BRFSS), seasonal temperature and humidity (weather), and number of hospital beds. We conduct analyses adjusting for additional variables: the total number of COVID-19 tests performed up to June 18, 2020, the estimated percent of people with COVID-19 symptoms and the longitude/latitude of each county. We repeat our analyses using four relevant sets of exposure data (P-1, P-2, P-3 and P-4).

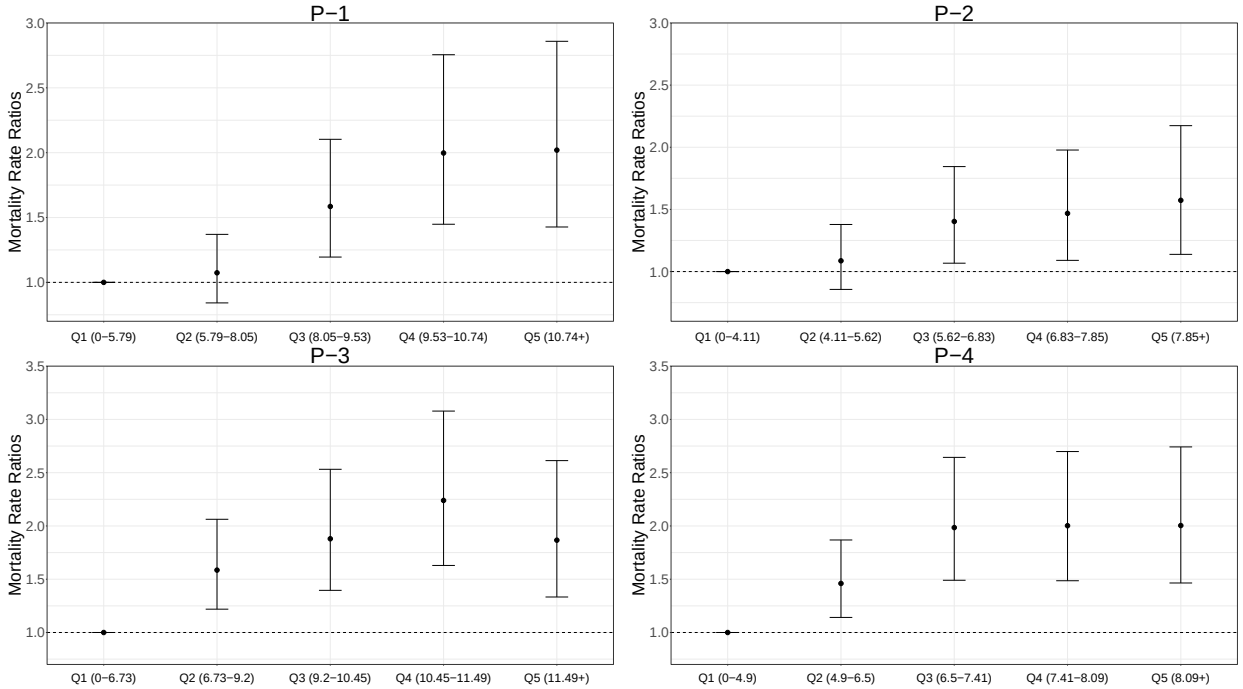

Figure S2: COVID-19 mortality rate ratios (MRR) per empirical quintile increase in $PM_{2.5}$ and 95% CI. The MRR can be interpreted as the percentage increase in the COVID-19 death rate associated with each empirical quintile increase of long-term average $PM_{2.5}$ compared to the baseline quintile (Q1). We repeat our analyses using four relevant sets of exposure data (P-1, P-2, P-3 and P-4).

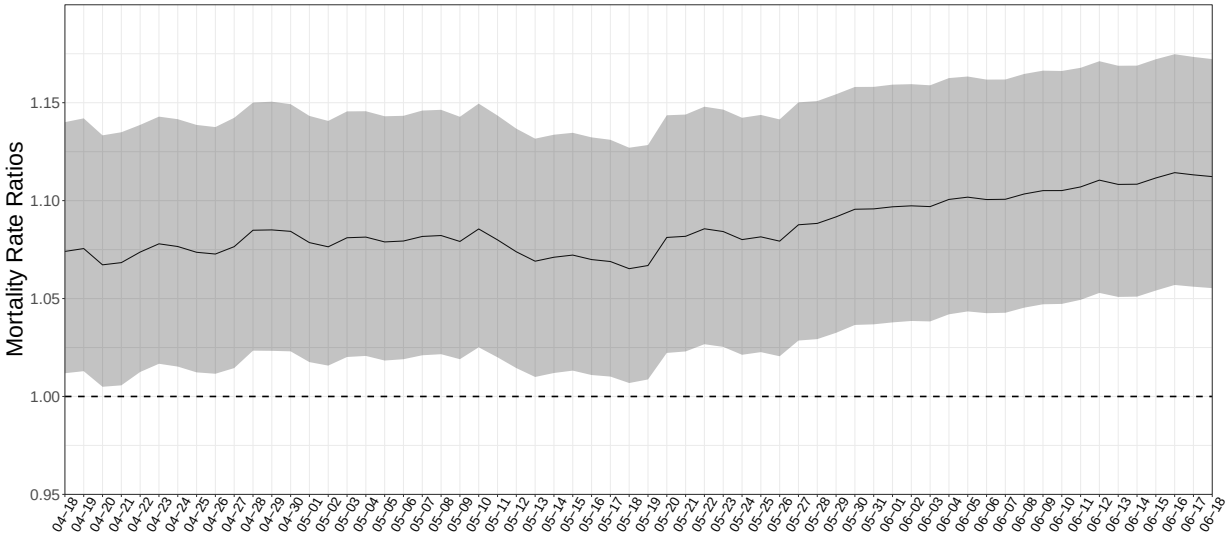

Figure S3: Daily COVID-19 mortality rate ratios (MRR) per 1 μg/m^3 increase in $PM_{2.5}$ and 95% CI. We conduct our main analysis using daily cumulative COVID-19 death counts from April 18, 2020 to June 18, 2020.

Table S1: Publicly available data sources used in the analyses.

	Source	Data
Outcome: COVID-19 Deaths	Johns Hopkins University the Center for Systems Science and Engineering (JHU-CSSE) Coronavirus Resource Center (https://coronavirus.jhu.edu/)	County-level COVID-19 death count up to and including June 18, 2020
Exposure: $PM_{2.5}$ concentrations	Atmospheric Composition Analysis Group (https://sites.wustl.edu/acag/)	$0.01^\circ \times 0.01^\circ$ grid resolution $PM_{2.5}$ prediction, averaged across the period 2000–2016 and averaged across grid cells in each county
Confounders for main analysis	US Census/American Community Survey (https://www.census.gov/programs-surveys/acs/data.html)	County-level socioeconomic and demographic variables for 2012–2016
	Robert Wood Johnson Foundation County Health Rankings (https://www.countyhealthrankings.org/)	County-level behavioral risk factor variables for 2020
	JHU-CSSE Coronavirus Resource Center	Time since first reported COVID-19 case
	Raifman et al., Boston University School of Public Health, COVID-19 United States state policy database (https://github.com/USCOVIDpolicy/COVID-19-US-State-Policy-Database)	Time since issuance of stay-at-home order
	Homeland Infrastructure Foundation-Level Data (HIFLD) (https://hifld-geoplatform.opendata.arcgis.com)	County-level number of hospital beds in 2019
	Gridmet via Google Earth engine (http://www.climatologylab.org/gridmet.html)	4 km $\times$ 4 km temperature and relative humidity predictions, summer and winter averaged across the period 2000–2016 and averaged across grid cells in each county
Additional confounders for secondary analyses	The COVID tracking project (https://covidtracking.com/)	State level number of COVID-19 tests performed up to and including June 18, 2020
	Carnegie Mellon University Delphi Research Center (https://github.com/cmu-delphi/delphi-epidata)	Estimated percentage of people with COVID-19 symptoms, based on survey data

Table S2: Characteristics of the study cohort up to and including June 18, 2020, mean (standard deviation).

	Total (3,087 counties)	$PM_{2.5}$< 8 $\mu g/m^3$ (1,217 counties)	$PM_{2.5}$≥ 8 $\mu g/m^3$ (1,870 counties)
COVID-19 death rate (per 100,000)	15.5 (31.2)	7.0 (19.7)	21.0 (35.7)
Average $PM_{2.5}$($\mu g/m^3$)	8.4 (2.5)	5.7 (1.4)	10.1 (1.2)
Rate of hospital beds (per 100,000)	300.8 (428.4)	338.8 (548.2)	276.1 (325.3)
Days since first case	76.9 (21.1)	68.2 (27.8)	82.6 (12.2)
Days since stay-at-home order	59.4 (38.1)	52.4 (41.6)	63.9 (34.9)
% Smokers	17.4 (3.5)	15.8 (3.1)	18.5 (3.4)
% Obese	32.9 (5.4)	31.2 (5.1)	34.0 (5.3)
% In poverty	10.5 (5.9)	9.6 (5.8)	11.0 (5.9)
% Less than high school education	21.3 (10.7)	16.6 (8.8)	24.4 (10.6)
% Owner-occupied housing	75.0 (8.3)	76.4 (7.6)	74.1 (8.6)
% Hispanic	7.5 (12.3)	9.6 (13.6)	6.1 (11.1)
% Black	8.0 (14.1)	1.0 (1.9)	12.6 (16.5)
% ≥65 years of age	16.0 (4.1)	17.4 (4.6)	15.0 (3.4)
% 45-64 years of age	26.4 (3.0)	26.9 (3.8)	26.1 (2.4)
% 15-44 years of age	37.6 (6.5)	35.2 (8.2)	39.2 (4.5)
Population density (person/km^2)	149.5 (699.2)	46.1 (146.9)	216.9 (884.4)
Median household income ($1,000)	49.3 (13.4)	50.6 (11.1)	48.4 (14.6)
Median house value ($1,000)	135.7 (89.9)	140.4 (87.8)	132.7 (91.1)
Average summer temperature (°F)	86.0 (5.7)	83.7 (6.7)	87.4 (4.4)
Average winter temperature (°F)	45.1 (11.9)	39.4 (11.4)	48.7 (10.7)
Average summer relative humidity (%)	89 (9.6)	83.2 (11.5)	92.8 (5.5)
Average winter relative humidity (%)	87.5 (4.8)	88.0 (5.6)	87.2 (4.1)

Table S3: Main, secondary and sensitivity analysis results for exposure data P-1, i.e., $PM_{2.5}$ exposure measured as the 17-year average concentration 2000-2016 by van Donekelaar et al. (9). Point estimates, 95 % confidence intervals, and p-values for the mortality rate ratio (MRR) for $PM_{2.5}$.

Analysis	N Counties	MRR (CI)	P-Value
Main analysis	3089	1.11(1.06, 1.17)	0.00
Omit # beds	3089	1.11(1.06, 1.17)	0.00
Omit BRFSS	3089	1.11(1.06, 1.17)	0.00
Omit weather	3089	1.12(1.07, 1.17)	0.00
Omit outbreak time	3089	1.13(1.07, 1.19)	0.00
Exclude counties in New York Metropolitan	3062	1.11(1.05, 1.17)	0.00
Exclude counties with <10 confirmed cases	2436	1.11(1.05, 1.17)	0.00
Rural counties	1940	1.07(0.98, 1.16)	0.14
Urban counties	1149	1.09(1.02, 1.16)	0.01
Categorize PM into quintiles	3089		
Q1 (0-5.79)		0	
Q2 (5.79-8.05)		1.07(0.84, 1.37)	0.57
Q3 (8.05-9.53)		1.59(1.02, 2.10)	0.00
Q4 (9.53-10.74)		2.00(1.45, 2.76)	0.00
Q5 (10.74+)		2.02(1.43, 2.86)	0.00
Add # tested	3089	1.11(1.05, 1.17)	0.00
Add COVID-like symptons	1468	1.14(1.08, 1.21)	0.00
Add Longitude and Latitude	3089	1.10(1.03, 1.18)	0.00
Add Longitude only	3089	1.09(1.03, 1.16)	0.00
Add Latitude only	3089	1.13(1.07, 1.19)	0.00
Adjust log(population density) as covariate	3089	1.11(1.05, 1.17)	0.00
Adjust log(population) as covariate	3089	1.12(1.06, 1.18)	0.00
Adjust population as covariate	3089	1.16(1.09, 1.23)	0.00
Use zero inflated Negative Binomial model	3089	1.11(1.06, 1.17)	0.00
Use fixed effects Negative Binomial model	3089	1.11(1.05, 1.17)	0.00

1. Five boroughs of New York City are considered as one county aligning with COVID-19 statistics.
2. "Rural" represents "Micropolitan" and "Non-core" counties defined by 2013 NCHS Urban-Rural Classification Scheme.

Table S4: Main, secondary and sensitivity analysis results for exposure data P-2, i.e., $PM_{2.5}$ exposure measured as 2016 average by van Donekelaar et al. (9). Point estimates, 95 % confidence intervals, and p-values for the mortality rate ratio (MRR) for $PM_{2.5}$.

Analysis	N counties	MRR (CI)	P-Value
Main analysis	3089	1.09(1.03, 1.16)	0.00
Omit # beds	3089	1.09(1.03, 1.16)	0.00
Omit BRFSS	3089	1.09(1.03, 1.16)	0.00
Omit weather	3089	1.11(1.05, 1.17)	0.00
Omit outbreak time	3089	1.11(1.05, 1.18)	0.00
Exclude counties in New York Metropolitan	3062	1.10(1.03, 1.16)	0.00
Exclude counties with <10 confirmed cases	2436	1.09(1.03, 1.16)	0.00
Rural counties	1940	1.02(0.93, 1.13)	0.68
Urban counties	1149	1.10(1.03, 1.18)	0.00
Categorize PM into quintiles	3089		
Q1 (0-4.11)		0	
Q2 (4.11-5.62)		1.09(0.86, 1.38)	0.49
Q3 (5.62-6.83)		1.40(1.07, 1.84)	0.02
Q4 (6.83-7.85)		1.47(1.09, 1.98)	0.01
Q5 (7.85+)		1.57(1.14, 2.17)	0.01
Add # tested	3089	1.09(1.03, 1.16)	0.00
Add COVID-like symptons	1468	1.15(1.08, 1.23)	0.00
Add Longitude and Latitude	3089	1.08(1.01, 1.15)	0.02
Add Longitude only	3089	1.07(1.01, 1.14)	0.03
Add Latitude only	3089	1.11(1.05, 1.17)	0.00
Adjust log(population density) as covariate	3089	1.08(1.01, 1.15)	0.02
Adjust log(population) as covariate	3089	1.09(1.03, 1.16)	0.00
Adjust population as covariate	3089	1.13(1.06, 1.21)	0.00
Use zero inflated Negative Binomial model	3089	1.09(1.03, 1.16)	0.00
Use fixed effects Negative Binomial model	3089	1.09(1.02, 1.15)	0.01

Table S5: Main, secondary and sensitivity analysis results for exposure data P-3, i.e., $PM_{2.5}$ exposure measured as the 17-year average concentrations 2000-2016 by Di et al. (29). Point estimates, 95 % confidence intervals, and p-values for the mortality rate ratio (MRR) for $PM_{2.5}$.

Analysis	N counties	MRR (CI)	P-Value
Main analysis	3089	1.11(1.05, 1.17)	0.00
Omit # beds	3089	1.11(1.05, 1.17)	0.00
Omit BRFSS	3089	1.10(1.05, 1.16)	0.00
Omit weather	3115	1.11(1.06, 1.17)	0.00
Omit outbreak time	3089	1.14(1.08, 1.20)	0.00
Exclude counties in New York Metropolitan	3062	1.11(1.05, 1.17)	0.00
Exclude counties with <10 confirmed cases	2436	1.09(1.03, 1.15)	0.00
Rural counties	1940	1.09(1.00, 1.18)	0.04
Urban counties	1149	1.07(1.01, 1.14)	0.02
Categorize PM into quintiles	3089		
Q1 (0-6.73)		0	
Q2 (6.73-9.2)		1.59(1.22, 2.06)	0.00
Q3 (9.2-10.45)		1.88(1.40, 2.53)	0.00
Q4 (10.45-11.49)		2.24(1.63, 3.08)	0.00
Q5 (11.49+)		1.87(1.33, 2.61)	0.00
Add # tested	3089	1.11(1.05, 1.16)	0.00
Add COVID-like symptons	1468	1.14(1.07, 1.21)	0.00
Add Longitude and Latitude	3089	1.12(1.05, 1.19)	0.00
Add Longitude only	3089	1.12(1.06, 1.19)	0.00
Add Latitude only	3089	1.11(1.06, 1.17)	0.00
Adjust log(population density) as covariate	3089	1.10(1.05, 1.16)	0.00
Adjust log(population) as covariate	3089	1.11(1.05, 1.17)	0.00
Adjust population as covariate	3089	1.13(1.07, 1.20)	0.00
Use zero inflated Negative Binomial model	3089	1.11(1.05, 1.17)	0.00
Use fixed effects Negative Binomial model	3089	1.10(1.05, 1.16)	0.00

Table S6: Main, secondary and sensitivity analysis results for exposure data P-4, i.e., $PM_{2.5}$ exposure measured as 2016 average by Di et al. (29). Point estimates, 95 % confidence intervals, and p-values for the mortality rate ratio (MRR) for $PM_{2.5}$.

Analysis	N counties	MRR (CI)	P-Value
Main analysis	3089	1.15(1.07, 1.23)	0.00
Omit # beds	3089	1.15(1.07, 1.23)	0.00
Omit BRFSS	3089	1.14(1.07, 1.23)	0.00
Omit weather	3115	1.14(1.07, 1.21)	0.00
Omit outbreak time	3089	1.17(1.09, 1.26)	0.00
Exclude counties in New York Metropolitan	3062	1.15(1.08, 1.24)	0.00
Exclude counties with <10 confirmed cases	2436	1.14(1.06, 1.22)	0.00
Rural counties	1940	1.09(0.97, 1.22)	0.13
Urban counties	1149	1.14(1.05, 1.24)	0.00
Categorize PM into quintiles	3089		
Q1 (0-4.9)		0	
Q2 (4.9-6.5)		1.46(1.14, 1.87)	0.00
Q3 (6.5-7.41)		1.98(1.49, 2.64)	0.00
Q4 (7.41-8.09)		2.00(1.49, 2.70)	0.00
Q5 (8.09+)		2.00(1.46, 2.74)	0.00
Add # tested	3089	1.15(1.07, 1.23)	0.00
Add COVID-like symptons	1468	1.20(1.11, 1.30)	0.00
Add Longitude and Latitude	3089	1.14(1.05, 1.22)	0.00
Add Longitude only	3089	1.17(1.09, 1.25)	0.00
Add Latitude only	3089	1.17(1.09, 1.25)	0.00
Adjust log(population density) as covariate	3089	1.13(1.06, 1.22)	0.00
Adjust log(population) as covariate	3089	1.15(1.07, 1.23)	0.00
Adjust population as covariate	3089	1.19(1.11, 1.29)	0.00
Use zero inflated Negative Binomial model	3089	1.15(1.07, 1.23)	0.00
Use fixed effects Negative Binomial model	3089	1.15(1.07, 1.23)	0.00