# Supplementary Materials for Exposure to air pollution and COVID-19 mortality in the United States

## Updated April 5 2020

## Contents




Authors: Xiao Wu M.S., Rachel C. Nethery Ph.D., M. Benjamin Sabath, M.A., Danielle Braun Ph.D., Francesca Dominici Ph.D.

# S.1 Data

## Health data

The Johns Hopkins University Center for Systems Science and Engineering created and maintains a platform hosting worldwide coronavirus case and death count data at the national and sub-national level that are updated in real time. For the US, these data are provided by the US Centers for Disease Control and Prevention (CDC) and state goverment at the county level. As of April 4, 2020, the CDC reports that COVID-19 testing is being conducted at 95 public health laboratories across the US and territories. To our knowledge, the CDC has not yet made publicly available information about how COVID-19 deaths are identified, i.e., whether a death attributed to COVID-19 requires a positive test or can be based on symptoms alone. Therefore, it remains unclear at this early stage how accurately COVID-19 death counts are being captured.

## Pollution data

We rely on modeled $PM_{2.5}$ exposure estimates rather than monitored observations alone, because air pollution monitors are sparsely distributed across the US, with a large majority of counties not containing a monitor. Our primary $PM_{2.5}$ modeled exposure estimates are produced by van Donekelaar et al (2019) (1). They are created 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 US, then this surface is bias-corrected for ground-monitor $PM_{2.5}$ observations using a geographically-weighted regression. The cross-validated $R^2$ for these models in the US was reported to be 0.61, although the accuracy varies across regions. For the primary analysis, the gridded data were averaged across the years 2000-2016 and then were aggregated to the county level using area-weighting. For sensitivity analyses, we also considered the 2016 county average $PM_{2.5}$, created using an

analogous procedure.

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 US using another well-validated ensemble-based prediction model (2). 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-weighting approach to aggregate the gridded data across the years 2000-2016 and then aggregate to the county level.

## Potential Confounders

To adjust for confounding bias in the nationwide observational study, we use county level variables from numerous public sources. Multiple socioeconomic and demographic variables were collected from the 2000 and 2010 Census (`https://www.census.gov`) and the 2005–2016 American Community Surveys (`https://www.census.gov/programs-surveys/acs/`). Specifically, we collect the following nine county level census variables: proportion of residents older than 65, proportion of Hispanic residents, proportion of Black residents, median household income, median home value, proportion of residents in poverty, proportion of residents with a high school diploma, population density, and proportion of residents that own their house. We also collect two county-level health risk factors from the Behavioral Risk Factor Surveillance System (BRFSS) (`https://www.cdc.gov/brfss/index.html`): average body mass index and smoking rate. We use the BRFSS variables from 2011, as this is the most recent year with county-level data available. 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 has been performed up to April 04, 2020 from the COVID tracking project (`https://covidtracking.com/`). We obtain meteorological variables on maximum daily temperature and relative humidity data on 4km x 4km gridded rasters from Gridmet via Google Earth Engine (`https://developers.google.com/earth-engine/datasets/catalog/IDAHO_EPSCOR_GRIDMET`). We average daily temperature and relative humidity for the summer (June-September) and winter (December-February) period respectively across the period 2000-2016 and average across grid rasters in each county. We also adjust for all four of these weatheer variables in our main models. The data used for this study are publicly available and sources are listed in Table 1 of the main manuscript.

## S.2 Statistical Modeling

We fit zero-inflated Negative Binomial regression models with a state-specific random intercept (3; 4). Zero-inflated negative binomial models are composed of two sub-models. The "count sub-model" estimates the association between the covariates and the number of COVID-19 deaths among counties eligible to experience a COVID-19 death. Letting $E[\cdot]$ denote an expected value, it takes the form

$$\log(\text{E[COVID-19 deaths]}) = \beta_0^c + \beta_1^c \text{ PM}_{2.5} + \beta_2^c \text{ population density} + \beta_3^c \text{ percent of the population older than 65-year old} + \beta_4^c \text{ percent living in poverty} + \beta_5^c \text{ median household income} + \beta_6^c \text{ percent black} + \beta_7^c \text{ percent hispanic} + \beta_8^c \text{ percent of the adult population with less than a high school education} + \beta_9^c \text{ median house value} + \beta_{10}^c \text{ percent of owner-occupied housing} + \beta_{11}^c \text{ average BMI} + \beta_{12}^c \text{ smoking rate} + \beta_{13}^c \text{ number of hospital beds} + \beta_{14}^c \text{ average summer temperature} + \beta_{15}^c \text{ average summer relative humidity} + \beta_{16}^c \text{ average winter temperature} + \beta_{17}^c \text{ average winter relative humidity} + \beta_{18}^c \text{ number of COVID-19 tests performed in state} + \log(\text{population}$$

size) + random intercept(State)

The "zero sub-model" accounts for the excess or structural zeros in the data that may be generated by counties not eligible for COVID-19 deaths, e.g., due to the absence of confirmed COVID-19 cases. It has the form

$$\text{logit}(\text{E}[\text{COVID-19 death ineligible}]) = \beta_0^z + \beta_1^z \text{ PM}_{2.5} + \beta_2^z \text{ population density} + \beta_3^z \text{ percent of the population older than 65-year old} + \beta_4^z \text{ percent living in poverty} + \beta_5^z \text{ median household income} + \beta_6^z \text{ percent black} + \beta_7^z \text{ percent hispanic} + \beta_8^z \text{ percent of the adult population with less than a high school education} + \beta_9^z \text{ median house value} + \beta_{10}^z \text{ percent of owner-occupied housing} + \beta_{11}^z \text{ average BMI} + \beta_{12}^z \text{ smoking rate} + \beta_{13}^z \text{ number of hospital beds} + \beta_{14}^z \text{ average summer temperature} + \beta_{15}^z \text{ average summer relative humidity} + \beta_{16}^z \text{ average winter temperature} + \beta_{17}^z \text{ average winter relative humidity} + \beta_{18}^z \text{ number of COVID-19 tests performed in state} + \text{random intercept(State)}$$

The $\beta^z$ estimates from the zero sub-model reflect the association between the covariates and a county's odds of being a structural zero, i.e., ineligible for a COVID-19 death. For the count sub-models, we provide the mortality rate ratios (MRR) and 95% CIs for $\text{PM}_{2.5}$, corresponding to the exponentiated parameter estimate ($e^{\hat{\beta}_1^c}$). The MRR can be interpreted as the multiplicative increase in the COVID-19 death rate associated with a 1 $\mu g/m^3$ increase in long-term average $\text{PM}_{2.5}$ exposure among the counties having the possibility to have COVID-19 deaths as of April 4, 2020. It is unclear whether the results of the zero sub-model provide any meaningful insights, as it is likely that the structural zeros in this setting arise due to the absence of the spread of COVID-19 in a community. It remains unclear whether $\text{PM}_{2.5}$ could be expected to impact the spread of COVID-19.

## Model Assumption Diagnostics

### Over-dispersion

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 ($\mu = 2.29$); however the variance is large due to the large death counts in several outbreak epicenters ($\sigma^2 = 1290.10$). Among the county with non-zero deaths, the mean of our outcome data is still relative small ($\mu = 10.30$); however the variance is large ($\sigma^2 = 5724.75$). The dispersion parameter for quasi-Poisson family taken to be 6.95, which indicate a strong 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.

### Zero-inflation

The total number of counties included in our main analysis is $3,080$, of which $2,395$ (77.8%) had not reported any COVID-19 deaths by April 4, 2020. We tested for the zero-inflation by plotting the expected outcome from a Negative Binomial regression with random effects vs. the observed outcome in the real data set. Figure S1 shows that there are substantially more zeros in the observed outcome compared to the expected outcome from a Negative Binomial regression with random effects. We anticipate that these zeros arise due to the absence of COVID-19 cases in some counties on/before April 4, 2020, making them ineligible to experience a COVID-19 death. For these reasons, we chose to fit zero-inflated negative binomial models to the COVID-19 death data.

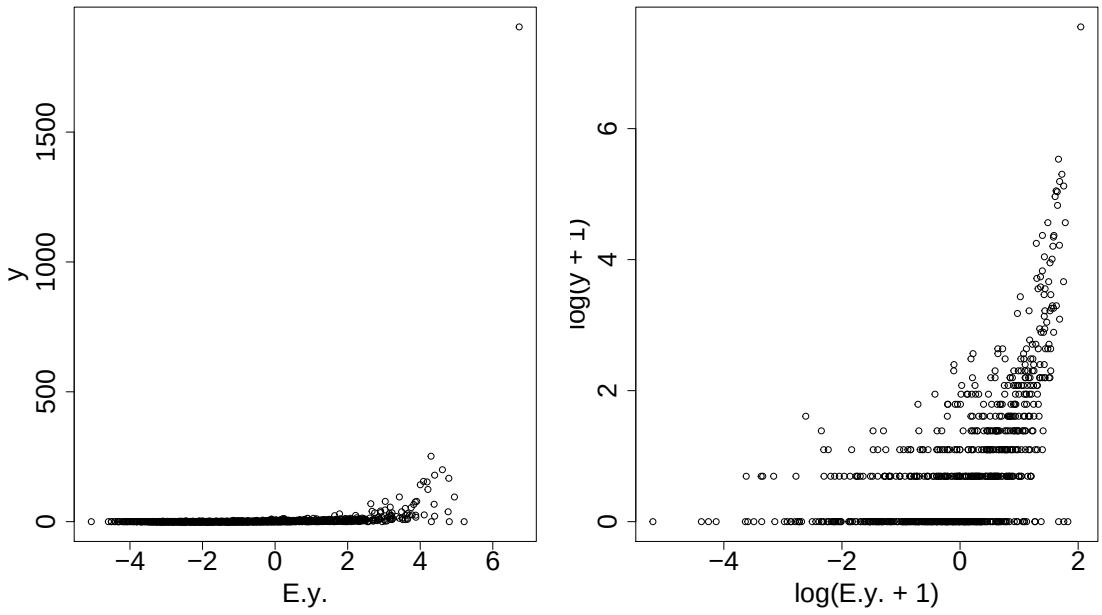


Figure S1: Diagnostic Plot for Zero-inflation. We compare the (log transformed) expected outcome from a Negative Binomial regression with random effects vs. the observed outcome in the real data set. Note if $Y$ is zero, then $\log(Y+1)$ will still be zero. There are substantially more zeros in the observed outcome compared to the expected outcome from a Negative Binomial regression with random effects.

## S.3 Additional Analysis Results

The detailed results are presented in Table S1-S4 and Figure S2. 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 (1), 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 (2019) (2), 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 the main text and in Section S1. 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 give similar point estimates for $PM_{2.5}$ and attain statistical significance using each of the different pollution data sources. 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.

As described in the main text, for each of these pollution data sources, we also evaluate the model sensitivity to the set of confounders adjusted for by individually omitting each of the following from the confounder set: 1) the number of hospital beds in the county; 2) the number of COVID-19 tests performed in each state; 3) behavioral risk factors, i.e., population mean BMI and percent of population who are smokers; and 4) meteorological (weather) variables: the summer (June-September) and winter (December-February) average of maximum daily temperatures and relative humidity in the county across 17 years (2000-2016). Effect estimates are presented as mortality rate ratio (MRR) per 1 $\mu g/m^3$increase in annual $PM_{2.5}$. We consistent positive associations between long-term exposure to $PM_{2.5}$ and increased mortality for COVID-19 in these analyses, with MRR between $1.06 - 1.15$ across

P-1 models that adjust for different potential confounders (similar results for P-2, P-3, and P-4). The removal of the number of hospital beds from the confounder set consistently attenuated the significance of the estimates, suggesting that the number of hospital beds is a strong confounder.

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 state 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 state, we still find a statistically significant association between long-term exposure to $PM_{2.5}$ and increased mortality for COVID-19 with MRR 1.13 and 95% confidence interval $(1.03, 1.22)$ for P-1. In the analysis that excludes counties with less than 10 confirmed COVID-19 cases, we also find a significant positive association with increased mortality of COVID-19 with magnitude of MRR 1.12 and 95% confidence interval $(1.01, 1.22)$ for P-1.

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 as a categorical variable (categorized at empirical quintiles), 3) using a negative binomial model without accounting for zero-inflation, 4) adjusting for population size as a covariate, rather than as an offset. In the analysis that treats $PM_{2.5}$ as a categorical variable, we found the magnitude of MRRs increase dramatically and monotonically as the quintile of $PM_{2.5}$ exposures increases for P-1. Such findings suggest there is no threshold about the effect of long-term exposure to $PM_{2.5}$ on COVID-19 mortality. In the analysis that adjusts for population density as a categorical variable (categorized at empirical quintiles), we again find a significant positive association with increased COVID-19 mortality with MRR 1.14 for P-1. In the analysis that uses a negative binomial model without accounting for zero-inflation, we find very similar results as of 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 death, 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.

## S.4 Code

```
library("dplyr")
library("MASS")
library(NBZIMM)

glmm.zinb.off = glmm.zinb(fixed = Deaths ~ mean_pm25 + scale(poverty)
+ scale(popdensity)  +scale(medianhousevalue) +scale(medhouseholdincome)
+ scale(pct_owner_occ)  + scale(hispanic) + scale(education)  +scale(pct_blk)
+ scale(older_pecent) + scale(beds) + scale(mean_bmi) + scale(smoke_rate)
+ scale(mean_summer_temp) + scale(mean_winter_temp) + scale(mean_summer_rm)
+ scale(mean_winter_rm) + scale(totalTestResults) + offset(log(population)),
                 random = ~ 1 | state, data = (aggregate_pm_census_cdc_test_beds))

glmm.zinb.beds = glmm.zinb(fixed = Deaths ~ mean_pm25 + scale(poverty)
+ scale(popdensity)  +scale(medianhousevalue) +scale(medhouseholdincome)
+ scale(pct_owner_occ)  + scale(hispanic) + scale(education)  +scale(pct_blk)
+ scale(older_pecent) + scale(mean_bmi) + scale(smoke_rate)
+ scale(mean_summer_temp) + scale(mean_winter_temp) + scale(mean_summer_rm)
+ scale(mean_winter_rm) + scale(totalTestResults) + offset(log(population)),
                 random = ~ 1 | state, data = (aggregate_pm_census_cdc_test_beds))

glmm.zinb.beds = glmm.zinb(fixed = Deaths ~ mean_pm25 + scale(poverty)
```

```
+ scale(popdensity)  +scale(medianhousevalue) +scale(medhouseholdincome)
+ scale(pct_owner_occ)  + scale(hispanic) + scale(education)  +scale(pct_blk)
+ scale(older_pecent)  + scale(beds) + scale(mean_bmi) + scale(smoke_rate)
+ scale(mean_summer_temp) + scale(mean_winter_temp) + scale(mean_summer_rm)
+ scale(mean_winter_rm) + offset(log(population)),
                 random = ~ 1 | state, data = (aggregate_pm_census_cdc_test_beds))

glmm.zinb.beds = glmm.zinb(fixed = Deaths ~ mean_pm25 + scale(poverty)
+ scale(popdensity)  +scale(medianhousevalue) +scale(medhouseholdincome)
+ scale(pct_owner_occ)  + scale(hispanic) + scale(education)  +scale(pct_blk)
+ scale(older_pecent)  + scale(beds)
+ scale(mean_summer_temp) + scale(mean_winter_temp) + scale(mean_summer_rm)
+ scale(mean_winter_rm) + scale(totalTestResults) + offset(log(population)),
                 random = ~ 1 | state, data = (aggregate_pm_census_cdc_test_beds))

glmm.zinb.beds = glmm.zinb(fixed = Deaths ~ mean_pm25 + scale(poverty)
+ scale(popdensity)  +scale(medianhousevalue) +scale(medhouseholdincome)
+ scale(pct_owner_occ)  + scale(hispanic) + scale(education)  +scale(pct_blk)
+ scale(older_pecent)  + scale(beds) + scale(mean_bmi) + scale(smoke_rate)
+ scale(totalTestResults) + offset(log(population)),
                 random = ~ 1 | state, data = (aggregate_pm_census_cdc_test_beds))
```

## S.5 Figures

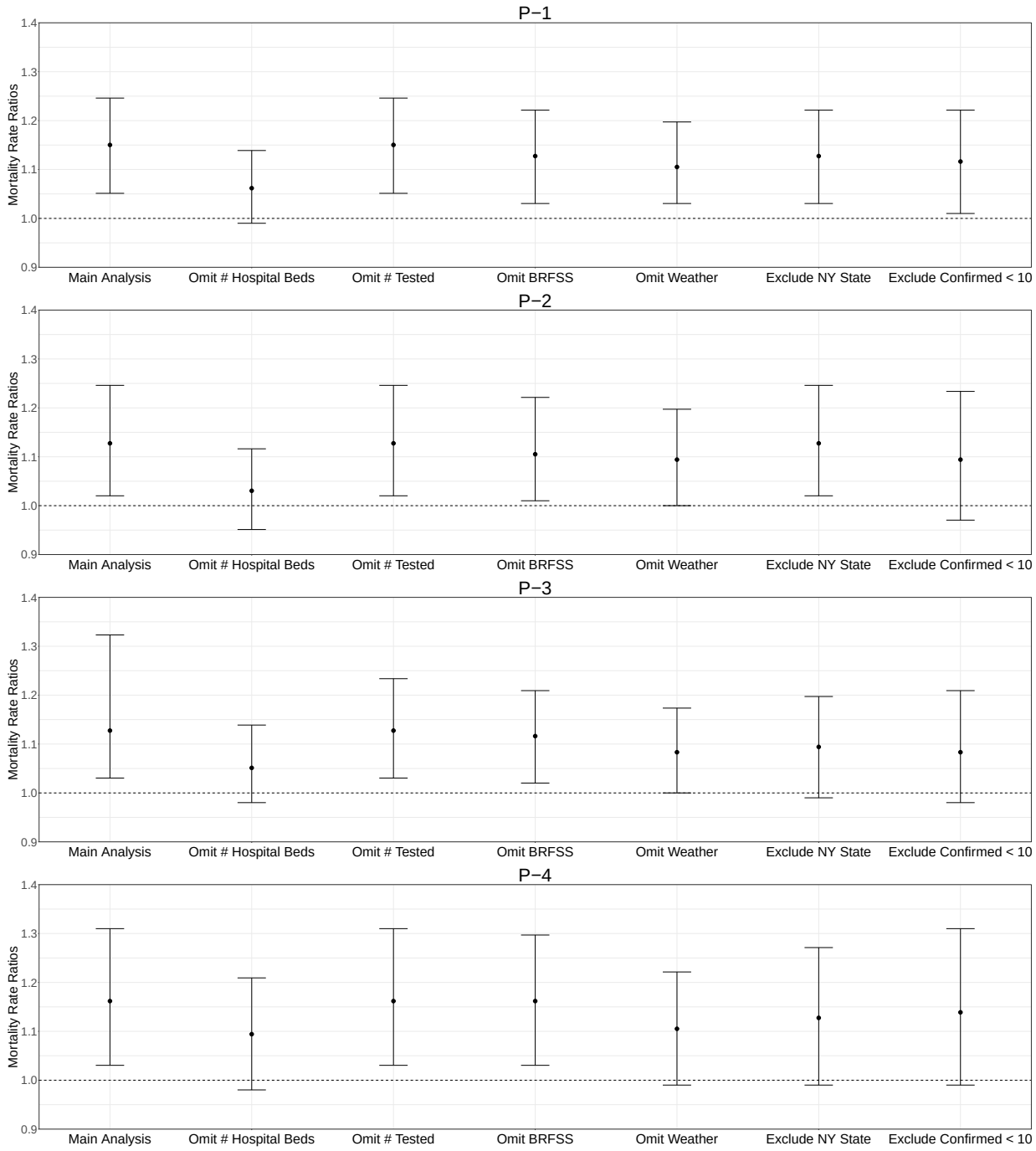


Figure S2: COVID-19 mortality rate ratios (MRR) per 1 $\mu$g/m$^3$ increase in $PM_{2.5}$ and 95% CI. The main analyses were adjusted for 17 socioeconomic, demographic, behavioral, climate, and healthcare confounders. We additionally conduct analyses omitting the following variables from the adjustment set: number of hospital beds, number of COVID-19 tests in each state, smoking rate and BMI from BRFSS, and seasonal temperature and humidity variables (weather). We also fit models excluding counties from NY state and excluding counties with $< 10$ confirmed cases. We repeat our analyses using four relevant sets of exposure data (P-1, P-2, P-3 and P-4).

## S.6 Tables

Table S1: Main, secondary and sensitivity analysis results for P-1, i.e., $PM_{2.5}$ exposure measured as the 17-year average concentration 2000-2016 by van Donekelaar et al (2019) (1). Point estimates and 95 % confidence intervals and p-values of the log mortality rate ratios (MRR).

| Analysis | $N$ Counties | log(MRR) | P-Value |
|---|---|---|---|
| Main analysis | 1783 counties | $0.14(0.05, 0.22)$ | $< 0.01$ |
| Exclude # beds | 2214 counties | $0.06(-0.01, 0.13)$ | 0.09 |
| Exclude # tested | 1783 counties | $0.14(0.05, 0.22)$ | $< 0.01$ |
| Exclude BRFSS | 2272 counties | $0.12(0.03, 0.20)$ | 0.01 |
| Exclude weather | 1783 counties | $0.10(0.03, 0.18)$ | 0.01 |
| Exclude counties in New York | 1726 counties | $0.12(0.03, 0.20)$ | 0.01 |
| Exclude counties with $< 10$ confirmed cases | 873 counties | $0.11(0.01, 0.20)$ | 0.04 |
| Categorize $PM_{2.5}$ into quintiles | 1783 counties | | |
| Q1 (0-5.79 $\mu g/m^3$) | | 0 | |
| Q2 (5.79-8.05 $\mu g/m^3$) | | $0.36(-0.13, 0.85)$ | 0.15 |
| Q3 (8.05-9.53 $\mu g/m^3$) | | $0.65(0.11, 1.19)$ | 0.02 |
| Q4 (9.53-10.74 $\mu g/m^3$) | | $0.89(0.31, 1.46)$ | $< 0.01$ |
| Q5 (10.74+ $\mu g/m^3$) | | $1.23(0.60, 1.85)$ | $< 0.01$ |
| Categorize population density into quintiles | 1783 counties | $0.13(0.03, 0.22)$ | 0.01 |
| Use standard Negative Binomial model | 1783 counties | $0.14(0.05, 0.23)$ | $< 0.01$ |
| Adjust log(population) as covariate | 1783 counties | $0.19(0.10, 0.28)$ | $< 0.01$ |
| Adjust population as covariate | 1783 counties | $0.37(0.28, 0.46)$ | $< 0.01$ |

Table S2: Main, secondary and sensitivity analysis results for P-2, i.e., $PM_{2.5}$ exposure measured as 2016 average by van Donekelaar et al (2019) (1). Point estimates and 95 % confidence intervals for the log mortality rate ratios (MRR).

| Analysis | $N$ counties | log(MRR) | P-Value |
|---|---|---|---|
| Main analysis | 1783 counties | $0.12(0.02, 0.22)$ | 0.02 |
| Exclude # beds | 2214 counties | $0.03(-0.05, 0.11)$ | 0.50 |
| Exclude # tested | 1783 counties | $0.12(0.02, 0.22)$ | 0.02 |
| Exclude BRFSS | 2227 counties | $0.10(0.01, 0.20)$ | 0.04 |
| Exclude weather | 1783 counties | $0.09(0.00, 0.18)$ | 0.06 |
| Exclude counties in New York | 1726 counties | $0.12(0.02, 0.22)$ | 0.02 |
| Exclude counties with $< 10$ confirmed cases | 716 counties | $0.09(-0.03, 0.21)$ | 0.12 |
| Categorize $PM_{2.5}$ into quintiles | 1783 counties | | |
| Q1 (0-4.11 $\mu g/m^3$) | | 0 | |
| Q2 (4.11-5.61 $\mu g/m^3$) | | $-0.46(-0.91, -0.01)$ | 0.04 |
| Q3 (5.61-6.82 $\mu g/m^3$) | | $0.00(-0.49, 0.49)$ | 1.00 |
| Q4 (6.82-7.85 $\mu g/m^3$) | | $0.12(-0.39, 0.64)$ | 0.65 |
| Q5 (7.85+ $\mu g/m^3$) | | $0.28(-0.27, 0.83)$ | 0.32 |
| Categorize population density into quintiles | 1783 counties | $0.10(0.00, 0.21)$ | 0.06 |
| Use standard Negative Binomial model | 1783 counties | $0.13(0.03, 0.23)$ | 0.01 |
| Adjust log(population) as covariate | 1783 counties | $0.17(0.07, 0.27)$ | $< 0.01$ |
| Adjust population as covariate | 1783 counties | $0.38(0.28, 0.48)$ | $< 0.01$ |

Table S3: Main, secondary and sensitivity analysis results for P-3, i.e., $PM_{2.5}$ exposure measured as the 17-year average concentrations 2000-2016 by Di et al (2019) (2). Point estimates and 95 % confidence intervals for the log mortality rate ratios (MRR).

| Analysis | $N$ counties | log(MRR) | P-Value |
|---|---|---|---|
| Main analysis | 1783 counties | $0.12(0.03, 0.28)$ | 0.01 |
| Adjust # beds | 2214 counties | $0.05(-0.02, 0.13)$ | 0.15 |
| Adjust # tested | 1783 counties | $0.12(0.03, 0.21)$ | 0.01 |
| Adjust BRFSS | 2272 counties | $0.11(0.02, 0.19)$ | 0.01 |
| Adjust weather | 1787 counties | $0.08(0.00, 0.16)$ | 0.04 |
| Exclude counties in New York | 1726 counties | $0.09(-0.01, 0.18)$ | 0.07 |
| Exclude counties with $< 10$ confirmed cases | 873 counties | $0.08(-0.02, 0.19)$ | 0.1 |
| Categorize $PM_{2.5}$ into quintiles | 1783 counties | | |
| Q1 (0-6.71 $\mu g/m^3$) | | 0 | |
| Q2 (6.71-9.19 $\mu g/m^3$) | | $0.37(-0.12, 0.85)$ | 0.13 |
| Q3 (9.19-10.45 $\mu g/m^3$) | | $0.50(-0.04, 1.04)$ | 0.07 |
| Q4 (10.45-11.48 $\mu g/m^3$) | | $0.50(-0.08, 1.07)$ | 0.09 |
| Q5 (11.48+ $\mu g/m^3$) | | $0.68(0.07, 1.28)$ | 0.03 |
| Categorize population density into quintiles | 1783 counties | $0.11(0.02, 0.20)$ | 0.01 |
| Use standard Negative Binomial model | 1783 counties | $0.13(0.03, 0.22)$ | 0.01 |
| Adjust log(population) as covariate | 1783 counties | $0.14(0.05, 0.23)$ | $< 0.01$ |
| Adjust population as covariate | 1783 counties | $0.27(0.18, 0.37)$ | $< 0.01$ |

Table S4: Main, secondary and sensitivity analysis results for P-4, i.e., $PM_{2.5}$ exposure measured as 2016 average by Di et al (2019) (2). Point estimates and 95 % confidence intervals for the log mortality rate ratios (MRR).

| Analysis | $N$ counties | log(MRR) | P-Value |
|---|---|---|---|
| Main analysis | 1783 counties | $0.15(0.03, 0.27)$ | 0.01 |
| Adjust # beds | 2214 counties | $0.09(-0.02, 0.19)$ | 0.10 |
| Adjust # tested | 1783 counties | $0.15(0.03, 0.27)$ | 0.02 |
| Adjust BRFSS | 2272 counties | $0.15(0.03, 0.26)$ | 0.01 |
| Adjust weather | 1787 counties | $0.10(-0.01, 0.20)$ | 0.07 |
| Exclude counties in New York | 1726 counties | $0.12(-0.01, 0.24)$ | 0.06 |
| Exclude counties with $< 10$ confirmed cases | 873 counties | $0.13(-0.01, 0.27)$ | 0.07 |
| Categorize $PM_{2.5}$ into quintiles | 1783 counties | | |
| Q1 (0-4.89 $\mu g/m^3$) | | 0 | |
| Q2 (4.89-6.50 $\mu g/m^3$) | | $0.37(-0.08, 0.83)$ | 0.11 |
| Q3 (6.50-7.41 $\mu g/m^3$) | | $0.60(0.07, 1.13)$ | 0.03 |
| Q4 (7.41-8.09 $\mu g/m^3$) | | $0.72(0.18, 1.26)$ | 0.01 |
| Q5 (8.09+ $\mu g/m^3$) | | $0.82(0.25, 1.39)$ | $< 0.01$ |
| Categorize population density into quintiles | 1783 counties | $0.14(0.01, 0.26)$ | 0.03 |
| Use standard Negative Binomial model | 1783 counties | $0.15(0.03, 0.28)$ | 0.02 |
| Adjust log(population) as covariate | 1783 counties | $0.17(0.05, 0.29)$ | $< 0.01$ |
| Adjust population as covariate | 1783 counties | $0.30(0.18, 0.42)$ | 0.01 |

# References


1. Van Donkelaar A, Martin RV, Li C, Burnett RT. Regional estimates of chemical composition of fine particulate matter using a combined geoscience-statistical method with information from satellites, models, and monitors. *Environmental science & technology.* 2019;53(5):2595–2611.

2. Di Q, Amini H, Shi L, *et al.* An ensemble-based model of PM2. 5 concentration across the contiguous United States with high spatiotemporal resolution. *Environment international.* 2019;130:104909.

3. Yau KK, Wang K, Lee AH. Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros. *Biometrical Journal: Journal of Mathematical Methods in Biosciences.* 2003;45(4):437–452.

4. Zhang X, Mallick H, Tang Z, *et al.* Negative binomial mixed models for analyzing microbiome count data. *BMC bioinformatics.* 2017;18(1):4.