# Supplementary Information

#

# S.1 Data and sources

Our paper uses the urban mortality rate derived from India’s Vital Statistical System that is generated through the Civil Registration System (CRS) and the Sample Registration Survey (SRS), two chief sources of vital records in India. The registration of deaths is carried out under provisions of the Registration of Births and Deaths (RBD) Act, 1969, which requires all deaths to be registered by the local registrar appointed by the state governments.

We collected data on registered number of deaths from the CRS for 110 districts (out of a total of 466 districts spread across 25 Indian states and 7 territories as per the 1991 Census of India) focusing on the urban areas for the years 1998 – 2015. Here, an urban district is an administrative division within an Indian state or territory that is a conglomeration of towns and cities. Using the Census population figures for these districts in the urban areas, we calculated the crude death rate (CDR). Note that a few district definitions have changed over the years and for the purposes of this research, we used the state and the district definition as per the 1991 Census.

Although the RBD Act of 1969 has been amended several times to improve the system of registration as well as to bring about uniformity across the country, there are still substantial differences between the actual number of deaths and registered deaths, with actual deaths being much higher than what is registered. The measure that determines how well these vital events are recorded is the level of registration, defined as the percentage of registered deaths out of the total deaths as estimated through the SRS for each state in India. The SRS uses a representative sample of India to estimate and project the state-level death rate, which is an estimate of the actual death rate. We derive the registration rate, which indicates the completeness of the data for all deaths in the country, by dividing the CDR by the estimated death rate from SRS. Because the registration rate for India is available only at the state level, we use this state-level registration and the district-level CDR obtained from CRS to calculate the estimated district-level mortality rate by assuming that the performance of the district’s urban areas in recording these vital events is the same for all districts within the state. This estimated urban mortality rate for each district that we derive from the above methodology is what we will from here on refer to as “adjusted mortality rate”.

The urban adjusted mortality rate (AMR), which is an average of 706 deaths per 100,000 people for this sample, is over 8 percent higher than the urban CDR from CRS (see Table S1 for summary statistics). This difference between the AMR and the CDR has declined over time showing that the system has become better at recording these vital events. The AMR has increased 9.4% while the CDR has increased 13.4% between 1998 and 2015, from which we can infer that the former statistic is more comprehensive than the one from the CRS. For these reasons, we will show the results from the model that uses AMR as the outcome variable.

**Table S1. Summary statistics**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Mean | Standard | 10th | Median | 90th |
|  |  | Deviation | Percentile |  | Percentile |
| *(Observations=1,727)* | **[2]** | **[3]** | **[4]** | **[5]** | **[6]** |
| **Panel A: Outcome:** |  |  |  |  |  |
| Urban Adjusted Mortality Rate | 706.3 | 545.6 | 415 | 667.2 | 972.3 |
| Urban Crude Death Rate | 651.1 | 242.7 | 379.2 | 645 | 925 |
| **Panel B: Pollution Measure:** |  |  |  |  |  |
| PM2.5 (µgm-3) | 49.7 | 26.9 | 25.3 | 40.7 | 97.2 |
| **Panel C: Control Variable:** |  |  |  |  |  |
| Urban Literacy Rate (%) | 78.5 | 6.5 | 69.9 | 80.8 | 85 |

Sources: The urban AMR are from CRS data that are adjusted using SRS estimates; the urban CDR are from CRS data from the Census in India; PM2.5 data are from satellite observations; literacy rates are from the Census in India.

Note: The Urban AMR and Urban CDR are rates per 100,000 persons.

The district-level data on PM2.5 concentrations were collected from the Atmospheric Composition Analysis Group 1 for the 18-year time period. The average PM2.5 concentration for 2015 is 58 µgm-3, a 39 percent increase from the 1998 concentrations.

In determining the impact of pollution on mortality, literacy is a covariate that affects the demand for better air quality as well better health. To control for this, we collected literacy rates for urban areas at the state level from the Census of India from three census years, 1991, 2001, and 2011. For the years in between, we interpolated the numbers to get the values for the 18 years in the dataset. Government data shows that urban literacy has increased from an average of 71 percent in 1998 across the districts we are interested in, to 79 percent in 2015.

# S.2 Model selection and goodness of fit

An important objective of the paper is to find the functional form of the relationship between AMR and PM2.5 as it relates to India. To understand this relationship, we started by using the Box-Cox transformation to inform us of the shape of this concentration-response (C-R) function. We used the methodology as suggested by Box and Cox 2, where the dependent variable and the main independent variable (here, PM2.5) were transformed by using different powers or logarithms. Using the maximum likelihood method, the estimate showed the log-log transformation fits the data best which has a “supralinear" shape.

To allow for greater flexibility in the specification and confirm, visually, what we found from the Box Cox transformation, we used a semi-parametric approach to understand the shape of the C-R function. Specifically, we ran a restricted cubic spline regression of the following form:

where is the adjusted mortality rate in district and year , is the PM2.5 for each district-year, and is the control variable we use, *i.e.*, urban literacy rate. represents the district level fixed effects and represents the year fixed effects. We have also included the district-specific linear and quadratic trends (denoted by and respectively) to control for any systematic trends specific to each district. The piecewise polynomial function, , or the spline, in Equation (S[1](#eq:eq8a)) allows the relationship to be nonlinear. We chose a model with three knots so that we can control for overfitting. For ease of understanding the shape identified from this exercise, we display the results graphically instead of the estimated coefficients. Figure S1 illustrates the C-R function has a supralinear shape, appearing similar to the log-log functional form suggested by the Box-Cox transformation method.



**Figure S1: Non-parametric spline regression.**

However, the spline estimation technique does not have an effective way of dealing with outliers in the dependent variable, like the Huber M-estimation model does. Despite the benefit of added flexibility in estimating the shape using the spline method, we ultimately decided to use the log-log functional form (*i.e.*, in Equation (1)) with Huber M-estimation as our preferred model to properly account for the large number of vertical outliers in the data.

We ran several specifications in addition to log-log and compared the goodness-of-fit measures of the models to the data. The additional models are log-linear (), log-linear squared (), and log-linear cubed (). The results of the model are displayed in Table S2.

Our results find that the log-log specification (Model [1] in Table S2) offers the best fit in terms of AIC and BIC measures. We used UCLA Statistical Consulting’s program, rregfit, to calculate the appropriate measures of fit for the models 3. Given the superior model fit of log-log over alternative specifications, the similarity of shape between the log-log and spline estimations, and the Box-Cox estimate indicating log-log is the appropriate shape, we chose the log-log model as our preferred specification.

**Table S2. Regression coefficients and measure of goodness of fit across estimated models.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | [1] | [2] | [3] | [4] |
| ln(PM2.5) | 0.104987\*\*\* |   |   |   |
|   | (0.040097) |   |   |   |
|   |   |   |   |   |
| PM2.5 |   | 0.001402\*\* | 0.004326\*\* | 0.004393 |
|   |   | (0.000695) | (0.001909) | (0.004497) |
|   |   |  |  |  |
| PM2.52 |   |  | 0.000020\* | -0.000021 |
|   |   |  | (0.000012) | (0.000063) |
|   |   |  |  |  |
| PM2.53 |   |  |  | 0.000000 |
|   |   |  |  | (0.000000) |
|   |   |   |   |   |
| AIC | 3,098.163 | 3,101.458 | 3,103.500 | 3,105.028 |
| BIC | 5,160.062 | 5,163.020 | 5,166.411 | 5,173.860 |
| R2 | 0.6804 | 0.6798 | 0.6804 | 0.6804 |

Notes: The dependent variable is the natural log of AMR. The number of observations is 1,727 district-years. All regressions are Huber M-estimation with standard errors in the parenthesis. All regressions include controls for urban literacy rate and district, year, district-specific linear and district-specific quadratic fixed effects (coefficients not shown in the table). The regressions are all population weighted. \*\*\*p<0.01, \*\*p<0.05, \*p<0.1.

# S.3 Identification Strategy

This section discusses all the identification issues related to endogeneity in estimating the C-R function between PM2.5 and mortality. These issues fall under three broad categories of endogeneity, that we elucidate in Sankar et al. (2020) 4: (i) reverse causality, (ii) unobserved heterogeneity, and (iii) measurement error.

## *Reverse Causality*

High pollution in certain areas could lead to sorting, where younger and healthier people move, leaving behind older and sicker people. This type of endogeneity would then bias the impact of pollution upwards, exaggerating the effect of pollution on health.

Our paper deals with this issue partly by using extensive fixed effects, controlling for district and year fixed effects, and also linear and quadratic trends within districts. Furthermore, Indian Census data indicates that dirty air is not the primary reason why people in India migrate within the country. As per the 2001 Indian Census, over 90 percent of migration within India is due to employment, marriage, education, business or familial reasons. Even if air pollution is a reason for some people in India to migrate, we believe it is small enough that reverse causality from mortality to ambient air pollution will not bias the results one way or the other.

A related question that could hinder causality is the effect of cumulative exposure on chronic illness, which could potentially magnify the impact of pollution. To test if this will bias the coefficient, we introduced lags into the model to see if pollution up to three years ago (following Lepeule et al. 5) can perhaps affect mortality in the current year. Tables [S3](#table:cumexp2) and S4 show that none of the lag terms are significant suggesting that this is not a concern for our sample of aggregate level annual data for India.

**Table S3. Effect of cumulative exposure of PM2.5 on mortality (log-log specification).**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|   | [1] | [2] | [3] | [4] |
| ln(PM2.5t) | 0.104987\*\*\* | 0.089948\*\* | 0.095702\*\* | 0.068102 |
|   | (0.040097) | (0.03911) | (0.041711) | (0.042187)  |
|   |   |   |   |   |
| ln(PM2.5t-1) |   | 0.0162 | 0.006846 | 0.043889 |
|   |   | (0.039841) | (0.041254) | (0.042243)  |
|   |   |   |   |   |
| ln(PM2.5t-2) |   |   | -0.038578 | -0.008229 |
|   |   |   | (0.041379) | (0.041100)  |
|   |   |   |   |   |
| ln(PM2.5t-3) |   |   |   | -0.025892 |
|   |   |   |   | (0.041551)  |
|   |   |   |   |   |
| Adjusted R2 | 0.680385 | 0.675587 | 0.684625 | 0.681921 |
| Observations  | 1,727 | 1,640 | 1,553 | 1,465 |

Notes: The dependent variable is the natural log of AMR. All results are from Huber M-estimations, and include controls for urban literacy rate and district, year, district-specific linear and district-specific quadratic fixed effects (coefficients not shown in the table). \*\*\*p<0.01, \*\*p<0.05, \*p<0.1.

**Table S4. Effect of cumulative exposure of PM2.5 on mortality (log-linear specification).**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|   | [1] | [2] | [3] | [4] |
| PM2.5t | 0.001402\*\* | 0.001218\* | 0.001716\*\* | 0.000692 |
|   | (0.000695) | (0.000661) | (0.000712) | (0.000760)  |
|   |   |   |   |   |
| PM2.5t-1 |   | 0.000516 | 0.000176 | 0.000287 |
|   |   | (0.000665) | (0.00068) | (0.000701)  |
|   |   |   |   |   |
| PM2.5t-2 |   |   | -0.000632 | -0.000331 |
|   |   |   | (0.00069) | (0.000665)  |
|   |   |   |   |   |
| PM2.5t-3 |   |   |   | -0.001386\*\* |
|   |   |   |   | (0.000701)  |
|   |   |   |   |   |
| Adjusted R2 | 0.679764 | 0.674622 | 0.685092 | 0.680512 |
| Observations  | 1,727 | 1,640 | 1,553 | 1,465 |

Notes: The dependent variable is the natural log of AMR. All results are from Huber M-estimations, and include controls for urban literacy rate and district, year, district-specific linear and district-specific quadratic fixed effects (coefficients not shown in the table). \*\*\*p<0.01, \*\*p<0.05, \*p<0.1.

In establishing causality, we also want to know if the cause comes strictly before the effect. If this is indeed true, then any lag or lead terms of pollution should have no effect on mortality in the current year. Tables S5 and S[6](#table:grangerll) show that the lag and lead terms do not have any effect on the mortality. This is true for both the log-log as well as the log-linear models.

**Table S5. Granger causality test of the effect of PM2.5 on mortality (log-log specification).**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|   | [1] | [2] | [3] | [4] |
| ln(PM2.5t-3) |   |   |   | 0.034222 |
|   |   |   |   | (0.059073)  |
|   |   |   |   |   |
| ln(PM2.5t-2) |   |   | -0.031477 | 0.045239 |
|   |   |   | (0.056948) | (0.074949)  |
|   |   |   |   |   |
| ln(PM2.5t-1) |   | -0.070979 | -0.128403\*\* | -0.019881 |
|   |   | (0.047158) | (0.05977) | (0.087575)  |
|   |   |   |   |   |
| ln(PM2.5t) | 0.104987\*\*\* | 0.054599 | 0.028738 | 0.074843 |
|   | -0.040097 | (0.050531) | (0.069918) | (0.102964)  |
|   |   |   |   |   |
| ln(PM2.5t+1) |   | -0.071923 | -0.102571 | -0.020849 |
|   |   | (0.055089) | (0.078574) | (0.105747)  |
|   |   |   |   |   |
| ln(PM2.5t+2) |   | 0.010052 | -0.012968 | 0.083438 |
|   |   | (0.053478) | (0.069754) | (0.095462)  |
|   |   |   |   |   |
| ln(PM2.5t+3) |   | -0.035711 | -0.02175 | 0.001228 |
|   |   | (0.052133) | (0.068391) | (0.088255)  |
|   |   |   |   |   |
| ln(PM2.5t+4) |   | -0.086263\* | -0.154347\*\* | -0.091153 |
|   |   | (0.049939) | (0.062018) | (0.072328)  |
|   |   |   |   |   |
| ln(PM2.5t+5) |   | -0.035399 | -0.073674 | -0.06014 |
|   |   | (0.045595) | (0.049777) | (0.055498)  |
|   |   |   |   |   |
| Observations  | 1,727 | 1,113 | 1,026 | 938 |
| Adjusted R2 | 0.967 | 0.985 | 0.985 | 0.987 |

Notes: The dependent variable is the natural log of AMR. All results are from Huber M-estimations, and include controls for urban literacy rate and district, year, district-specific linear and district-specific quadratic fixed effects (coefficients not shown in the table). \*\*\*p<0.01, \*\*p<0.05, \*p<0.1.

**Table S6. Granger causality test of the effect of PM2.5 on mortality (log-linear specification).**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|   | [1] | [2] | [3] | [4] |
| PM2.5t-3 |   |   |   | 0.001513 |
|   |   |   |   | (0.001142)  |
|   |   |   |   |   |
| PM2.5t-2 |   |   | 0.000422 | 0.001188 |
|   |   |   | (0.001166) | (0.001510)  |
|   |   |   |   |   |
| PM2.5t-1 |   | -0.000556 | -0.000561 | 0.000676 |
|   |   | (0.000861) | (0.001148) | (0.001765)  |
|   |   |   |   |   |
| PM2.5t | 0.001402\*\* | 0.001273 | 0.001771 | 0.002223 |
|   | (0.000695) | (0.000928) | (0.001446) | (0.002190)  |
|   |   |   |   |   |
| PM2.5t+1 |   | -0.000592 | -0.000077 | 0.000269 |
|   |   | (0.001091) | (0.001772) | (0.002406)  |
|   |   |   |   |   |
| PM2.5t+2 |   | 0.000948 | 0.001203 | 0.001761 |
|   |   | (0.001067) | (0.001606) | (0.002176)  |
|   |   |   |   |   |
| PM2.5t+3 |   | 0.000026 | 0.000586 | -0.000183 |
|   |   | (0.00109) | (0.001574) | (0.001994)  |
|   |   |   |   |   |
| PM2.5t+4 |   | -0.000336 | -0.001277 | -0.001543 |
|   |   | (0.001046 | (0.00132) | (0.001530)  |
|   |   |   |   |   |
| PM2.5t+5 |   | 0.000394 | 0.000141 | -0.001055 |
|   |   | (0.000897) | (0.00098) | (0.001040)  |
|   |   |   |   |   |
| Observations  | 1,727 | 1,113 | 1,026 | 938 |
| Adjusted R2 | 0.966 | 0.985 | 0.985 | 0.988 |

Notes: The dependent variable is the natural log of AMR. All results are from Huber M-estimations, and include controls for urban literacy rate and district, year, district-specific linear and district-specific quadratic fixed effects (coefficients not shown in the table). \*\*\*p<0.01, \*\*p<0.05, \*p<0.1.

## *Unobserved heterogeneity*

We use district-specific and year-specific fixed effects to remove most of the unobserved heterogeneity in the district-year observations. To account for possible systematic district and year-specific trends we include district-specific linear and quadratic time trends.

## *Measurement error*

As stated above, the CDR from the CRS is underreported. To correct for this, we use the registration rate in urban areas, as estimated by the SRS, to adjust the CDR and create the AMR to make it closer to the actual death rate. If the AMR is lower than the actual mortality rate, then this will lead to biasing the impact of pollution on mortality downwards. To deal with the possible measurement error in the AMR estimates, we use Huber M-estimation method, which is recommended for vertical outliers 6.

We do not expect any systematic measurement errors with respect to the PM2.5 data because they are estimated using satellite-based data as well as calibrated using surface PM2.5.

## *Violation of SUTVA*

Another confounding factor in identifying the effect of air pollution on mortality is the violation of SUTVA, *i.e.*, the effect of ambient air pollution in a certain district-year might affect the mortality rate in another district-year.

There are different ways in which this violation might have an effect. Within a district, air pollution in one year might affect mortality the following year. We tested for cumulative exposure and granger causality, as shown in Tables S3 – S6. Furthermore, pollution in a district in a certain year might also affect a person from a neighboring district that same year or even the following year. For instance, several people live in the neighboring areas surrounding the district of Delhi and can be affected by the dirty Delhi air that might travel to their area. This phenomenon could potentially affect their health in the current time period or the next. However, because the data are annual, there is less chance that pollution in the current year would affect health in the following year.

This effect is mostly accounted for by the battery of fixed effects that controls for any year effects affecting the whole nation or any district specific effects and trends. However, even if this is still an issue, we will need more structure that must be imposed on the model in order to deal with it. That would, along with all the fixed effects, introduce a lot more stress on the data. We understand that this is a limitation of this paper, but since the dataset has only 1,727 rows, we choose not include any more variables into this model.

## *Omitted Variable Bias*

Even after accounting for unobserved heterogeneity and other endogeneity issues through a battery of fixed effects, there could bias due omitted variables. To address this, we considered urban literacy rate and income. In our initial runs of the models, we included income as an indicator of economic activity, which was not significant and was removed from our subsequent regression runs. We include literacy rate to indicate the area’s economic activity as well as the role it might play in enhancing health conditions, similar to Pope et al. (1995) 7 and Pope et al. (2002) 8.

# S.4 Relative-risk equations and calculation of lives saved from reduction

The relative risk shows the ratio of the risk of mortality for a given concentration to the risk of mortality at a reference concentration. For our illustrations of the relative risk of mortality in Figures 1 and S1, the reference level is at the population-weighted average PM2.5 concentration equal to 49.7 µgm-3. Below we show the calculation of the relative risk for our two specifications log-log and log-linear (see Burnett et al. 9for the relative-risk equation for GEMM).

Define the PM2.5 concentration at the reference level as , and the concentration at any other location, , as . For the log-log and log-linear C-R functions the relative risk equations are

where the estimated parameters are and .

For our concentration-reduction scenarios we calculate the number of lives saved from a reduction in the PM2.5 concentration below the baseline level in each district. The baseline concentration, , is equal to the 2011 – 2015 average PM2.5 concentration for district . For these calculations, we use a modified relative-risk equation, where instead of setting the reference level at the average concentration, (as in Equation (S2)), the reference is equal to . With this modification, the relative risk is less than one for any concentration below the baseline.

Define the baseline mortality rate and population in district (average from 2011 – 2015) as and , respectively. The mortality rate for a concentration, , below the baseline is . The change in the mortality rate from the baseline is . And the number of lives saved () from a concentration reduction is the change in the mortality rate times the population:

which can be calculated for any of the relative-risk equations: log-log, log-linear and GEMM.

**S.5 Concentration-reduction scenarios**

In the *Equal* scenario, the PM2.5 concentration in every district is reduced by 10 µgm-3: . The budget of person-PM2.5 units, which is fixed for all three scenarios, is .

In the *Standard* scenario, a limit on PM2.5, , is set, and all districts with are reduced to the limit. The person-PM2.5 units used for any given limit is . The limit selected in the *Standard* scenario is the one in which the person-PM2.5 units used is less than the budget, with the fewest units remaining:

The level that satisfies this expression is µgm-3.

In the *Optimized* scenario, the concentration reductions are prioritized according to their ability to maximally reduce the mortality rate. This process, through a series of iterations (identified by ), calculates the marginal change in the mortality rate in every district, chooses the district with the maximum change, and reduces the concentration by a small amount: µgm-3. The person-PM2.5 units spent from this reduction, made in district for iteration , are calcuated as . The process is repeated (with the updated PM2.5 concentration in district ) until the budget is exhausted. If the concentration in any district reaches 5 µgm-3 it is removed from the process for all future iterations.

The marginal change in the mortality rate (from a concentration reduction) is:

For the log-log and log-linear relative-risk equations, the marginal relative risk is

**S.5 Sensitivity analysis**

*Log-log combined with GEMM for low PM2.5 concentrations*

Our concentration reduction scenarios requires reducing PM2.5 concentrations to levels below our lowest observed concentration in the dataset (17.1 µgm-3). As a sensitivity analysis, we present here the results from our Figure 2 combining the log-log and GEMM C-R functions. For concentrations in our dataset (PM2.5 ≥ 17.1) we use the log-log C-R function, and for concentrations below our lowest observed concentration we use the GEMM C-R function. This most importantly affects our *Optimized* scenario. Figure S2 illustrates the lives saved/exposure inequality relationship for this combined function in orange, and in lighter orange we reproduce the results from the original log-log C-R function. The combined function results in 12,800 fewer lives saved (10% fewer) in the *Optimized* scenario than using log-log.



**Figure S2. Sensitivity analysis combining log-log with GEMM for low PM2.5 concentrations.** The results are recalculated using the log-log C-R function for PM2.5 concentrations in the observed range (PM2.5 ≥ 17.1 µgm-3), and the GEMM C-R function is used for concentrations below the lowest observed concentration (PM2.5 < 17.1 µgm-3).

# References

1. van Donkelaar, A. *et al.* Global Estimates of Fine Particulate Matter using a Combined Geophysical-Statistical Method with Information from Satellites, Models, and Monitors. *Environ. Sci. Technol.* **50**, 3762–3772 (2016).

2. Box, G. E. P. & Cox, D. R. An analysis of transformations. *J. R. Stat. Soc.* **B 26**, 211–252 (1964).

3. Bruin, J. newtest: command to compute new test. *UCLA: Statistical Consulting Group* https://stats.idre.ucla.edu/stata/ado/analysis/ (2011).

4. Sankar, A., Coggins, J. S. & Goodkind, A. L. Effectiveness of air pollution standards in reducing mortality in India. *Resour. Energy Econ.* **62**, (2020).

5. Lepeule, J., Laden, F., Dockery, D. & Schwartz, J. Chronic Exposure to Fine Particles and Mortality: An Extended Follow-up of the Harvard Six Cities Study from 1974 to 2009. *Environ. Health Perspect.* **120**, 965–970 (2012).

6. Verardi, V. & Croux, C. Robust regression in Stata. *Stata J.* **9**, 439–453 (2009).

7. Pope III, A. C. *et al.* Particulate Air Pollution as a Predictor of Mortality in a Prospective Study of U.S. Adults. *Am. J. Respir. Crit. Care Med.* **151**, 669–674 (1995).

8. Pope III, A. C. *et al.* Lung Cancer, Cardiopulmonary Mortality, and Longterm Exposure to Fine Particulate Air Pollution. *J. Am. Med. Assoc.* **287**, 1132–1141 (2002).

9. Burnett, R. *et al.* Global estimates of mortality associated with longterm exposure to outdoor fine particulate matter. *Proc. Natl. Acad. Sci.* **115**, 9592–9597 (2018).