The Mortality and Medical Costs of Air Pollution: Evidence from Changes in Wind Direction

By Tatyana Deryugina, Garth Heutel, Nolan H. Miller, David Molitor, and Julian Reif

ONLINE APPENDIX

June 2019

Table of Contents

A. Source of identifying variation

We estimate the effect of pollution on mortality and health utilization over a broad geographic area without requiring a detailed "case study" of each individual location. In this section, we illustrate the variation in pollution that drives our results and explicitly test for concerns that may arise when using variation in pollution from an unspecified source.

To illustrate the relationship between wind and pollution for each of our monitor groups, we first estimate the following regression separately for each of the 100 monitor groups described in the main text:

$$\text{PM2.5}_{cdmy} = \sum_{b=0}^{34} \beta_b WINDDIR_{cdmy}^{10b} + f\left(Temp_{cdmy}, Prcp_{cdmy}, WindSpeed_{cdmy}\right) + \alpha_c + \alpha_{sm} + \alpha_{my} + \epsilon_{cdmy}. \quad \text{(A1)}$$

The variables $Temp_{cdmy}$ correspond to temperature bins, while $Prcp_{cdmy}$ and WS_{cdmy} correspond to precipitation and wind speed deciles, respectively, following the definitions in equations (1) and (2) of the main text. The function $f(\cdot)$ represents all their possible interactions. Likewise, the fixed effects follow equations (1) and (2). Equation (A1) differs from the first stage of the instrumental variable regressions estimated in the main text in three minor ways: (1) to demonstrate the source of our variation in more detail, it employs 10-degree bins for $WINDDIR$ instead of 90-degree bins; (2) because we are interested in the relationship between daily wind direction and daily fine particulate matter, this specification excludes leads and lags[1]; and (3) it does not employ county-population weights. Including more wind direction bins in our main analysis is computationally burdensome due to the large number of additional regressors it generates, though we do perform a robustness check to demonstrate the invariance of our results to more wind direction bins in Table 9.

Online Appendix Figure A1 plots estimates of $\hat{\beta}_b$ (solid black lines), along with their corresponding 95 percent confidence intervals (shaded grey area), for each of the monitor groups in our main sample.[2] The San Francisco Bay Area in California ("Santa Clara, CA") and the Boston, MA area ("Middlesex, MA") are reproduced in Figures 2 and 3 in the main text. For most monitor groups, there is a strong and clear relationship between local wind direction and changes in PM 2.5.

Our empirical approach raises the concern that a small number of monitors located close to local pollution sources may drive our first-stage results, while monitors located far from those sources may

[1] Leads were included in our main specification because our outcome variable there was three-day mortality and lags were included to minimize concerns about autocorrelation.

[2] The large number of control variables included in equation (A1) causes estimation to be impossible for seven of the 100 monitor groups (see notes in Online Appendix Figure A1).

exhibit no significant relationship between wind direction and pollution. If this is the case, then our estimates of the effect of pollution will be driven by local sources near pollution monitors, resulting in potentially significant measurement error. Because we do not observe pollution sources, we cannot test for this directly. However, we can provide indirect evidence by testing for the presence of outliers and by investigating whether the patterns from our first stage are similar for monitor groups located close together. To that end, we conduct two tests.

In the first test, we split each of our monitor groups into two random subgroups. We then estimate equation (A1) separately for each subgroup and compare the subgroup estimates to each other and to the group average. Intuitively, if a handful of monitors located near local sources drive our first stage, then the two subgroups should generate different results. By contrast, if the estimated patterns are driven by non-local transport, then the coefficients should be similar. As Figure A1 shows, for the vast majority of monitor groups, the coefficients for each of the two subgroups (dashed red lines) are qualitatively and quantitatively similar to each other and to the overall group average (solid black line), suggesting that our first stage is not driven by locally-produced pollution measured by a handful of nearby monitors.

In the second test, we first classify monitors into 50 groups instead of 100, using the same classification algorithm (kmeans).[3] We then match each monitor group from the 50-group classification to *all* overlapping groups from the 100-group classification. That is, for each of the 50 groups, we find all monitor groups in the 100-group set that have at least one monitor in common. Each of the 50 groups overlaps with 3 to 7 groups from the 100-group classification. We expect to see similar patterns in the first-stage estimates because these overlapping groups are located close to each other geographically, and air pollution can be carried by the wind for hundreds of miles.

Online Appendix Figure A2 shows the estimated wind angle-pollution relationship in each of the 50 groups (solid black line) and the corresponding relationships in the overlapping groups from the 100-group classification (dashed red lines). Intuitively, if the estimated patterns are driven by non-local transport, then the estimated coefficients should be qualitatively and quantitatively similar. Indeed, this is what we see in the vast majority of cases.

B. Medicare sample and mortality data

The baseline sample used in our analysis includes all Medicare beneficiaries ages 65–100 and is derived from 100 percent Medicare enrollment information files for years 1999–2013. These annual files include an observation for each beneficiary enrolled in Medicare for at least one day in that calendar year,

[3] We have also replicated our main results with 50 monitor groups (see column (2) of Table 9 in the main text), and they are very similar.

whether enrolled in Traditional Medicare (fee-for-service) or Medicare Advantage. The enrollment files report a variety of demographic and enrollment variables, including unique beneficiary identifiers that can be used to link individuals over time; monthly indicators for Medicare eligibility; state, county, and ZIP code of residence based on the mailing address for official correspondence; and exact dates of birth and death.

The vast majority of elderly living in the United States are enrolled in Medicare. The left panel of Online Appendix Figure A3 compares the size of our baseline Medicare sample to Census estimates of the US population ages 65 and over. To aid comparison, we use Census estimates of the resident population on July 1 each year and limit the Medicare sample to beneficiaries who reside in the 50 states and the District of Columbia and who turned 65 before July 1. Over the period 1999–2013, the Census estimates an average of 38.1 million elderly individuals each year, compared to 37.7 million elderly in Medicare. Thus, the Medicare sample covers over 98 percent of elderly living in the US, a share which remains roughly constant over the sample period.

The mortality variables used in our analysis are based on dates of death recorded in the Medicare enrollment files. Medicare's death data come primarily from the Social Security Administration but are augmented based on reviews triggered by hospitalization claims indicating patient death. The annual mortality rates in the Medicare data align closely with mortality rates based on National Vital Statistics death records and Census population estimates, as shown in the right panel of Online Appendix Figure A3. While all recorded deaths in the Medicare data are validated, some death *dates* in the data are not validated, in which case they are assigned the last date in the month of death. Because our analysis is performed at the daily level, we drop individuals who die at any point in the year and who do not have a validated death date flag. This restriction affects less than 2 percent of the deaths in our sample, and the share of deaths with unvalidated dates diminishes over time (see Online Appendix Figure A3).

To estimate PM 2.5 effects by cause of death, we use data from the National Death Index (NDI) created by the Center for Disease Control and matched to Medicare beneficiaries who died in 1999–2008. The NDI is a centralized database of death record information compiled from state vital statistics offices and maintained by the National Center for Health Statistics (NCHS).[4]

We first categorize ICD-10 cause of death codes into 39 groups based on the NCHS's list of 39 selected causes of death.[5] We then group these 39 causes of death into four categories as follows:

---

[4] For more information about the NDI, see https://www.cdc.gov/nchs/ndi.htm (accessed January 30, 2019).

[5] The list of 39 selected causes of death and the ranges of ICD-10 codes that comprise each cause are available at https://www.cdc.gov/nchs/data/dvs/im9_2002.pdf.pdf [*sic*] (accessed January 30, 2019).

1. **Cardiovascular deaths**: hypertensive heart disease with or without renal disease, ischemic heart disease, other diseases of the heart, essential (primary) hypertension and hypertensive renal disease, cerebrovascular diseases, atherosclerosis, other diseases of circulatory system
2. **Cancer deaths**: stomach cancer, colon cancer, pancreatic cancer, lung cancer, breast cancer, ovarian and uterine cancer, prostate cancer, bladder cancer, non-Hodgkin's lymphoma, leukemia, other cancer
3. **Other internal causes of death**: tuberculosis; syphilis; HIV; diabetes; Alzheimer's disease; influenza and pneumonia; chronic lower respiratory disease; peptic ulcer; chronic liver disease and cirrhosis; nephritis; pregnancy, childbirth and the puerperium; perinatal conditions; congenital abnormalities; SIDS; abnormal clinical findings; all other diseases
4. **External causes of death**: motor vehicle accidents, suicide, homicide, other accidents, other external causes

C. Predicting life expectancy

We predict life expectancy using two different survival models: Cox proportional hazards, and Survival Random Forest.[6] We estimate these models using individuals from our baseline sample who were alive and eligible for Medicare in 2002 and then use the estimates from these models to predict life expectancy for the other beneficiaries in our sample. To ensure that we have accurate measures of beneficiaries' baseline chronic conditions prior to 2002, we further limit the sample to Medicare beneficiaries who, as of January 1, 2002, had been continuously enrolled in fee-for-service Medicare for at least two years (and are thus at least 67 years old in 2002). We observe all deaths that occur among this cohort on or before December 31, 2013. During this 12-year time period (2002–2013), over 50 percent of our sample dies; the remaining deaths are censored.[7] For computational ease, we further limit the analysis to a random 5 percent sample of these beneficiaries. The final estimation sample used in our survival analysis includes 1,210,659 individuals.

C.1 Cox proportional hazards

The semi-parametric Cox proportional hazards model assumes that the hazard rate of death for individual i can be factored into two separate functions:

$$h(t_i|x_i,\beta) = h_0(t_i)\exp[x_i'\beta]$$

[6] We also estimated fully parametric models that assume survival rates are governed by either the Gompertz or Weibull distributions. Those yielded results similar to those from our preferred, less parametric, specifications.

[7] Although earlier cohorts are observable for a longer period of time, we do not use them because the Medicare variables denoting the presence of pre-existing chronic conditions, which are strong predictors of survival, are nonexistent or unreliable in earlier years.

The hazard rate at time t_i, $h(t_i|x_i, \beta)$, depends on the baseline hazard rate, $h_0(t_i)$, and on a vector of individual characteristics, x_i. The parameter vector β is estimated by maximizing the log partial likelihood function:

$$\ln L(\beta) = \sum_{i=1}^{N} \delta_i \left[x_i'\beta - \ln \sum_{j \in R(t_i)} \exp[x_j'\beta] \right], \tag{A2}$$

where the indicator variable δ_i is equal to one for individuals whose deaths we observe (uncensored observations) and equal to zero otherwise. The risk set $R(t_k) = \{l: t_l \geq t_k\}$ is the set of observations at risk of death at time t_k and consists of all individuals who are alive at that time. Thus, individuals whose deaths we do not observe (censored observations) affect the partial likelihood function only through the terms indexed by j in equation (A2).

Once $\hat{\beta}$ has been obtained by maximizing the log partial likelihood, we nonparametrically estimate the baseline hazard function following Breslow (1972):

$$\hat{h}_0(t_i) = \frac{d_{t_i}}{\sum_{j \in R(t_i)} \exp[x_j'\hat{\beta}]} \tag{A3}$$

The numerator, d_{t_i}, is the number of deaths that occur at t_i. The corresponding baseline survival function is calculated as

$$\hat{S}_0(t_i) = \exp[-\hat{H}_0(t_i)]$$

where $\hat{H}_0(t_i)$ is the cumulative hazard function, calculated as $\hat{H}_0(t_i) = \sum_{\tau=1}^{t_i} \hat{h}_0(\tau)$. The individual-specific survival function, which allows us to calculate life expectancy, can then be estimated as:

$$\hat{S}(t_i|x_i, \hat{\beta}) = \hat{S}_0(t_i)^{\exp[x_i'\hat{\beta}]}$$

In practice, the nonparametric estimate of the baseline hazard function is limited to the 12 years of Medicare data we have available for this survival analysis. We extrapolate the baseline hazard function to future years by assuming it follows a log-linear form. As shown in Online Appendix Figure A4, this appears to be a reasonable assumption.

The life-years lost analysis presented in the main text varies the set of individual characteristics included in the vector x_i in order to understand how they affect the results (see Table 4). As described in the text, we first estimate a standard Cox proportional hazards model using increasingly large sets of characteristics. The most detailed model that does not use machine learning includes age, sex, and indicators for 27 chronic conditions. We then turn to a specification that incorporates information from 1,062 variables. Including so many control variables creates two challenges. First, some variables may be significant predictors of survival for the 2002 cohort just by chance, even if they are not good predictors of survival in general. This may cause bias due to overfitting (Harrell, Lee, and Mark 1996). Second,

computational limitations prevent us from including a large set of regressors when performing conventional maximum likelihood estimation on a large sample using standard numerical procedures.

We overcome these challenges by estimating a Cox-Lasso model (Tibshirani 1997). Cox-Lasso can be implemented by maximizing a penalized version of objective function (A2):

$$\ln L(\beta) = \left(\sum_{i=1}^{N} \delta_i \left[x_i'\beta - \ln \sum_{j \in R(t_i)} \exp[x_j'\beta] \right] \right) - \lambda \sum_{i=1}^{k} |\beta_i| \qquad \text{(A4)}$$

where $|\beta_i|$ is the absolute value of β_i, the ith element of the vector β, and k is the number of included regressors. We select the optimal penalty parameter λ using five-fold cross validation.[8] We include the following 1,062 regressors (not including omitted categories) when estimating this survival model[9]:

1. Age in days as of January 1, 2002
2. Indicator variables for sex and for seven different races
3. Indicator variables for the presence of the following 27 different chronic conditions as of December 31, 2001: acute myocardial infarction, Alzheimer's disease, senile dementia, atrial fibrillation, cataracts, chronic kidney disease, chronic obstructive pulmonary disease (COPD), heart failure, diabetes, glaucoma, hip/pelvic fracture, ischemic heart disease, depression, osteoporosis, rheumatoid arthritis, stroke, breast cancer, colorectal cancer, prostate cancer, lung cancer, endometrial cancer, anemia, asthma, hyperlipidemia, benign prostatic hyperplasia, hypertension, and hypothyroidism
    a. Indicator variables for all pairwise interactions of these 27 chronic conditions
4. Indicator variables for the interaction of 27 chronic conditions with seven race indicators
5. Indicator variables for the interaction of 27 chronic conditions with sex
6. Indicator variables for 12 percentiles (10, 20, 30, 40, 50, 60, 70, 80, 90, 95, 99, 99.9) of the *beneficiary's* prior year spending (i.e., spending that excludes payments made by Medicare)
    a. Indicator variables for the same 12 quantiles for each of the following 17 different categories of *total* (beneficiary plus Medicare) prior year medical spending: hospice, home health care, hospital outpatient, acute inpatient, other inpatient, skilled nursing facility, ambulatory surgery center, Part B drugs, evaluation and management, anesthesia, dialysis, other procedures, imaging, tests, durable medical equipment, other Part B carrier, and Part B physician

[8] See Simon et al. (2011) for a detailed discussion of the algorithm we employ to implement the Cox proportional hazards estimator with a Lasso penalty term.

[9] These variables are described in detail in the ResDAC documentation: http://www.resdac.org/cms-data/files/mbsf-base.

7. Indicator variables for various percentiles (listed in parentheses) of the 2001 total annual number of:
    a. Dialysis events (10, 30, 50, 70, 90)
    b. Home health visits, hospital outpatient emergency room visits (10, 30, 50, 70, 90, 95)
    c. Anesthesia events, hospital outpatient visits, other Part B carrier events, acute inpatient stays, durable medical equipment (10, 30, 50, 70, 90, 99)
    d. Part B drug events (10, 50, 70, 90, 99, 99.5)
    e. Other procedures events, evaluation and management events, imaging events, hospital outpatient emergency room visits, tests events, Part B physician events (10, 30, 50, 70, 90, 99, 99.5)
8. Fourth-order polynomials in each of 37 different variables that have been merged to the respondent's five-digit ZIP code of residence. All variables are standardized so that they follow a normal distribution with mean zero and variance one. These ZIP code-level data are obtained from the 2007–2011 and 2008–2012 American Community Surveys. The variables include data on the following categories (number of variables in parentheses if more than one): travel time to work (2), fraction below the poverty line (3), median household income, aggregate household income, aggregate household social security income, aggregate household retirement income, fraction in labor force, heating fuel sources (3), aggregate number of vehicles, median home value, fraction immigrant, Gini index of household income, fraction with less than high school education, median year housing built, fraction on disability (2), fraction with hearing difficulties (2), fractions with vision difficulty (2), fraction with cognitive difficulty (2), fraction with ambulatory difficulty (2), fraction with self-care difficulty (2), fraction with independent-living difficulty (2), fraction with any health coverage (2), and fraction with private health coverage (2).

The estimated life expectancy that forms the basis of the estimate in Column (6) of Table 4 is based on estimating equation (A4) when including the 1,062 regressors listed above.

The dashed lines in Online Appendix Figure A5 show the distribution of estimated life expectancies for the subsample of Medicare beneficiaries used to estimate our survival model. The range of the distribution is wider when the model includes all 1,062 predictors (the dashed black line) than when it includes only age and sex as predictors (the dashed red line). The model based on age and sex corresponds to a typical life table that includes only 68 $(= (100 - 67 + 1) \times 2)$ values. (The maximum and minimum values in this life table correspond to life expectancies for a 67-year-old female and a 100-year-old male, respectively.) By contrast, the Cox-Lasso model generates a much larger set of predictions, some of which lie outside the range of a basic age-sex life table.

The solid lines in Online Appendix Figure A5 show how the distribution of predicted values changes when it is limited to the subset of beneficiaries who died during the 2002 calendar year. The distribution produced by the model that includes only age and sex—given by the solid red line—shifts to the left because these decedents are older than the average Medicare beneficiary and thus have below-average life expectancies. The distribution for the Cox-Lasso model—given by the solid black line—shifts to the left even more. This indicates that beneficiaries who died within one year of January 1, 2002 were not only older than the average beneficiary in that year, but also they were less healthy than average, as captured by variables like prior medical spending and prior chronic conditions. Accounting for these additional variables reduces (on average) the predicted life expectancies for these Medicare beneficiaries. This demonstrates that the Cox-Lasso model that incorporates data from many variables generates predictions that are more accurate than a simple Cox model that accounts only for age and sex.

To further validate these estimates, we perform a similar exercise that incorporates Medicare data from individuals not included in our estimation sample. We first use the estimates from our model to predict life expectancy for Medicare beneficiaries as of January 1 of each calendar year. For each of these years, we then calculate the average life expectancy for all fee-for-service beneficiaries who die during that year ("decedents"). We focus on this group because these decedents form the basis of the life-years lost estimates reported in Table 4.

Online Appendix Figure A6 displays the results of this exercise. The solid orange line, which serves as a baseline, displays our estimate of the unconditional life expectancy (11.4 years) for all Medicare beneficiaries. The solid green line displays life expectancy among "decedents," as predicted by a Cox proportional hazards model that conditions on age and sex. Because the typical decedent is older than the average beneficiary, the predictions from this model are about 3.5 years lower than the baseline. This is clearly a more accurate prediction, since these decedents by definition died within one year of when their life expectancy was estimated. For the sake of comparison, we also include predictions based on a period life table published by the Social Security Administration (SSA). The SSA life table conditions on age and sex, and its predictions are nearly identical to those of our Cox model estimated using age and sex. The solid black line displays estimates based on the Cox-Lasso estimation of the Cox proportional hazards model with 1,062 regressors. This reduces the prediction by yet another three years. The estimates decline slightly over time, which likely reflects the improvement in the recording of chronic conditions in later years.[10]

[10] Overall rates of chronic conditions captured in the Medicare data increase systematically each year from 1999 (the first year Medicare measured these conditions) to 2006. Because some chronic conditions are not treated regularly (and thus not diagnosed), these trends likely reflect incomplete measurement. Because beneficiaries in these earlier

C.2 Survival random forest

Random forest is a non-parametric, nonlinear method that predicts an outcome based on a set of inputs (Breiman 2001). The basic unit of a random forest is a decision tree. Decision trees sequentially split the predictor space into a number of simple regions. The predicted outcome for a given observation is typically the mean outcome among training observations that lie in the same predictor region.

A decision tree is grown using recursive binary splitting of the prediction space using a set of training observations. To begin, all training observations are contained in a single node. At each step of the recursive process, terminal nodes from the prior step are split by selecting the predictor and cut point that minimizes in-sample prediction error (e.g., residual sum of squares) of the resulting tree. Recursive binary splitting continues until a stopping rule is reached, resulting in a set of terminal nodes that each contain at least one observation. When there are a large number of variables, each with a large number of possible values, the decision tree can become very large.

While decision trees are simple and easy to interpret, they often have poor prediction accuracy. Random forest employs bootstrapping to generate many decision trees. The random forest prediction for an observation is calculated as the average prediction across all trees in the forest. The consensus prediction of a random forest can be much more accurate than the prediction from a single decision tree.

Ishwaran et al. (2008) extend random forest to cover right-censored survival data. We estimate a survival random forest model using the following algorithm (Ishwaran and Kogalur 2007):

1. For $b = 1$ to B trees:
    A. Draw a bootstrap sample of size N, where N is the number of observations in the dataset.
    B. Grow a decision tree until a minimum node size of s is reached, where minimum node size is defined as the number of deaths in that node.
        i. Select $m = \sqrt{k}$ variables at random, where k is the number of variables in the dataset.
        ii. Find the best split point among those m variables.
            a. The number of potential split points depends on the number of unique values realized by these m variables. For example, a binary variable like sex has one potential split point.

---

years are less likely to have their chronic conditions recorded in the data, their estimated life expectancy is higher than beneficiaries in later years, who are more likely to have recorded chronic conditions.

1) For each variable, limit the set of unique values to a random sample of $nsplit$ values. For example, if total spending has 1,000 different possible values and $nsplit = 10$, then a random sample of 10 spending values will be considered as potential split points.

b. The best split point is defined as the one that maximizes the value of the log-rank test statistic.

iii. Split the node into two daughter nodes, and then continue splitting subject to the constraint that a terminal node should not have less than s deaths.

2. Calculate the cumulative hazard function for each terminal node of each tree. Then, generate the prediction by averaging over all B trees.

A. The cumulative hazard function, $\widehat{H}_{b_i}(t)$ is equal to the number of deaths in node i of tree b at time t divided by the number of people at risk in that node at time t.

B. Let $\boldsymbol{I}[j \in \text{node } b_i]$ be an indicator function equal to 1 if individual i is a member of terminal node j in decision tree b and 0 otherwise. The predicted cumulative hazard for individual i is then obtained by averaging over all the terminal nodes in which she resides:

$$\widehat{H}(t|x_i) = \frac{1}{B}\sum_{b=1}^{B}\sum_{j}\boldsymbol{I}\big[i \in \text{node } b_j\big]\widehat{H}_{b_j}(t)$$

The survival function, $\hat{S}(t|x_i)$, can then be derived from the cumulative hazard function, $\widehat{H}(t|x_i)$, as described above. In our analysis, we set the number of trees B equal to 250, the minimum node size s equal to 3, and $nsplit$ equal to 10. As with the Cox proportional hazards model, the survival predictions are limited to the 12 years (2002–2013) for which we observe the 2002 Medicare cohort, so we again extrapolate the survival function to future years by assuming it follows a log-linear form.

We estimate our Survival Random Forest model at the monthly level using the same predictors as the Cox-Lasso model described above (age, sex, race, chronic conditions, medical spending, and ZIP-code-level demographics). Because Survival Random Forest inherently accounts for interactions and nonlinear effects, it requires fewer input variables (124) than Cox-Lasso (1,062). For example, we do not need to create indicator variables for different quantiles of spending, nor do we need to create pairwise interactions.

The dotted blue line in Online Appendix Figure A6 compares life expectancy estimates from Survival Random Forest to estimates from different Cox proportional hazards models. Unsurprisingly, Survival Random Forest using all available data performs significantly better than a Cox model that

includes only age and sex as predictors. However, it performs slightly worse than the Cox-Lasso model that also employs all available data.

D. Using machine learning to estimate heterogeneous treatment effects

Online Appendix Section D.1 describes the method developed by Chernozhukov, Demirer, Duflo, and Fernandez-Val (2018) (hereafter CDDF), as outlined in their paper. Online Appendix Section D.2 describes challenges in applying this method to our setting, and explains how we address these challenges.

D.1 Chernozhukov, Demirer, Duflo, and Fernandez-Val (2018)

CDDF develop a method for estimating heterogeneous treatment effects in randomized experiments that is valid even in high-dimensional settings. The setting outlined in their paper is as follows. Let Y be the outcome of interest and Z be a vector of covariates. Units are randomly assigned to either a treatment group ($T = 1$) or a control group ($T = 0$). The probability of assignment to treatment is given by the propensity score, $p(Z)$. Treatment effect heterogeneity is measured using the conditional average treatment effect function:

$$s_0(Z) = E[Y|T = 1, Z] - E[Y|T = 0, Z]$$

CDDF propose the following steps to study properties of $s_0(Z)$:

1. Split the sample into two approximately equal parts: a "main" and an "auxiliary" sample.
2. Use the auxiliary sample to train a machine learning (ML) algorithm (e.g., Lasso, Random Forest) to predict Y using Z. This prediction exercise is performed twice: once using only control group observations, and once using only treatment group observations.
3. Predict Y for observations *in the main sample* using both prediction models from step 2. That is, for each observation in the main sample, predict Y using estimates obtained from training the ML algorithm on the treated observations ($\widehat{Y_i^{T=1}}$), and predict outcomes using estimates obtained from training the ML algorithm on the control observations ($\widehat{Y_i^{T=0}}$).
4. Calculate the difference between these two predictions, $\hat{S}(Z) = \widehat{Y_i^{T=1}} - \widehat{Y_i^{T=0}}$.

The proxy predictor $\hat{S}(Z)$ is a possibly biased and inconsistent estimate of the conditional average treatment effect function, $s_0(Z)$. Nevertheless, CDDF show that the researcher can still use $\hat{S}(Z)$ to extract important *properties* of $s_0(Z)$. First, the researcher can identify $BLP\left[s_0(Z)|\hat{S}(Z)\right]$, the best linear predictor of $s_0(Z)$ using $S(Z)$, by estimating the following weighted regression:

$$Y = \beta_1\big(T - p(Z)\big) + \beta_2\big(T - p(Z)\big)\left(\hat{S}(Z) - \overline{\hat{S}}\right) + \theta' X_1 + \varepsilon, \tag{A5}$$

where the weights are equal to

$$w(Z) = \frac{1}{p(Z)\big(1 - p(Z)\big)}.$$

The control variables X_1 can include the baseline outcome prediction, $\widehat{Y_i^{T=0}}$, as well as a constant term.[11] The variable $p(Z)$ is the probability of treatment as a function of Z, i.e., the propensity score, which CDDF assume is known. Equation (A5) identifies the best linear predictor in the sense that $\beta_1 = E(s_0(Z))$ and $\beta_2 = Cov(s_0(Z), \hat{S}(Z))/Var(\hat{S}(Z))$. Testing whether $\beta_2 = 0$ corresponds to testing the joint null hypothesis of no heterogeneous treatment effects and irrelevance of the proxy predictor. In addition, CDDF suggest an alternative specification that employs the Horvitz-Thompson transformation:

$$YH = \beta_1 + \beta_2\left(\hat{S}(Z) - \overline{\hat{S}}\right) + \theta_H' X_1 H + \epsilon \tag{A6}$$

where

$$H = \frac{T - p(Z)}{p(Z)\big(1 - p(Z)\big)}.$$

Next, CDDF show how to obtain "sorted group average treatment effects." First, the researcher partitions the sample into K groups, $G_1, G_2, \ldots, G_K$, according to non-overlapping intervals of the proxy predictor, $\hat{S}(Z)$. To obtain the sorted effects, the researcher estimates the following weighted regression:

$$Y = \alpha' X_1 + \sum_{k=1}^{K} \gamma_k\big(T - p(Z)\big) \cdot 1(G_k) + \epsilon, \tag{A7}$$

where $1(G_k)$ is an indicator for belonging to group k and the weights, $w(Z)$, are the same ones used to estimate the best linear predictor of $s_0(Z)$. CDDF show that the estimated coefficient $\hat{\gamma}_k$ captures the average treatment effect in group k, $E(s_0(Z)|G_k)\ \forall k$. These coefficients can therefore be used to measure the distribution of treatment effects across these K groups.

Finally, CDDF also show that the average characteristics of units assigned to one of the K groups described above can be obtained by simply computing the mean. The researcher can then estimate differences in characteristics across these groups by comparing their means.

---

[11] As recommended by CDDF, we include $\widehat{Died^C}(Z_{it})$ as a control variable when estimating this and other regressions.

The estimated parameters above, such as $\hat{\gamma}_k$, are subject to two forms of uncertainty: sample splitting uncertainty from step 1 and standard estimation uncertainty. CDDF show that under sufficient regularity conditions the estimated parameters are normally distributed, conditional on the sample split in step 1. They therefore propose that the researcher repeat steps 1–4 and re-estimate equations (A5)–(A7) 100 times and then report the median of those 100 estimates, along with the medians of the 100 lower and upper bounds of the corresponding confidence intervals.

An appealing feature of CDDF is that the researcher can employ any ML algorithm when performing predictions. We use gradient boosted decision trees, as implemented by Chen and Guestrin (2016). Prior work has shown this algorithm to be a good predictor of mortality in the Medicare population (Einav et al. 2018). Our implementation allows for 500 boosting iterations and a maximum tree depth of 10.

D.2 Using CDDF to estimate mortality heterogeneity in a large-sample, non-RCT setting

We conduct our heterogeneity analysis using a person-day sample of beneficiaries who have been continuously enrolled in FFS for at least two years. Employing a disaggregated version of our county-day sample maximizes the amount of heterogeneity available for our analysis, but also introduces computational challenges because the disaggregation increases our sample size by a factor of 26,000. To minimize the computational load, our heterogeneity analysis employs a one-day window for the outcome instead of a three-day.

The CDDF method was designed for a randomized controlled trial (RCT) and thus requires clearly identified treatment and control groups. By contrast, our quasi-experimental setting has a continuous-valued endogenous regressor (pollution) and many binary instruments (wind direction bin indicators interacted with monitor group indicators). In order to apply CDDF to our setting, we assign each county-day observation in our sample to either a high pollution wind direction ("treatment") group or low pollution wind direction ("control") group as follows. We first estimate equation (A1) separately for each monitor group, as discussed in Section A of this Online Appendix. We then categorize wind direction bins with above-median estimated coefficients as "high pollution" directions, and those with below-median coefficients as "low pollution" directions. This allows us to define a new county-day indicator that is equal to one if the observed wind direction is associated with high pollution and zero otherwise. Online Appendix Table A12 reports estimates of the effect of PM 2.5 on one-day mortality using this just-identified IV specification. The specification presented in column (1), which includes the controls and fixed effects from our main empirical specification, estimates that a one-unit increase in PM 2.5 increases one-day mortality by 0.356 per million beneficiaries. For comparison, the corresponding over-identified IV specification

estimates a one-day mortality increase of 0.40 (not reported).[12] Columns (2)-(5) show how the just-identified estimate varies across different fixed-effect specifications.

Interpreting these estimates requires understanding how PM 2.5 levels differ across observations assigned to the treatment and control groups. Online Appendix Table A13 reports estimates of the first stage and the reduced form for a county-day specification estimated using the binary instrument described above. On average, PM 2.5 levels are 2.4 $\mu g/m^3$ higher for treated county-days than for control county-days (Panel A, columns (2)-(4)).

Another consequence of our non-RCT setting is that the propensity score is unknown. We therefore perform step 2 of the CDDF procedure twice: first to estimate mortality, Y, and second to estimate the propensity score, $p(Z)$.[13] Our analysis ignores the estimation error in the propensity score. This is reasonable because our sample size is enormous. To avoid overfitting, we do not allow the same people to appear in the main and auxiliary dataset when performing these two prediction exercises. In other words, the random assignment of observations to the main and auxiliary files is done at the person level, rather than the person-day level.

A second challenge we face in our setting is that the probability of an individual dying on any given day is very small. It is well-known that standard machine learning algorithms perform poorly in such cases because the algorithm will generally never predict death for anybody, which causes problems during step 2 of the CDDF procedure. We therefore follow Einav et al. (2018) and employ "downsampling" when we train our machine learning algorithm to predict mortality. Specifically, we sample every person-day in the auxiliary sample with a death outcome with probability one, and then randomly select the same number of auxiliary sample person-days with no death outcome, resulting in a perfectly balanced sample. We set aside 10 percent of this sample for calibrating the predicted death probability, as described next, and train the machine learning models on the remaining 90 percent.[14]

Because the death rate in this subsample is exactly 50 percent, the resulting predictions are biased upward and need to be adjusted. We correct for this bias in the same manner as Einav et al. (2018), using our calibration dataset and a Bayesian correction formula. First, we use the downsampled model to predict

---

[12] The corresponding over-identified IV estimate using three-day mortality is equal to 0.685 (see column (1) of Table 2 in the main text).

[13] The mortality estimation is performed using downsampled data, as described below. The propensity score estimation is performed using a random subsample of 24.5 million observations (0.1 percent of the auxiliary sample).

[14] There are 1.84 million person-days with a death outcome in the auxiliary control sample, including the 10 percent used for calibration. There are 2.07 million such person-days in the auxiliary treated sample.

mortality in the calibration dataset. We then regress the realized mortality rates in this 10 percent balanced sample on a cubic polynomial of the predicted mortality rates. This regression model is then used to adjust the mortality predictions for the observations in the main sample. We further adjust those predictions, $\hat{Y}^*$, using the following Bayesian correction formula:

$$\hat{Y} = \frac{\hat{Y}^*}{R - (R-1)\hat{Y}^*}$$

The value R is the ratio of survivors to decedents in the auxiliary sample. Incorporating this adjustment helps ensure that the final mortality predictions, $\hat{Y}$, match the mean death rate in the full sample.

A third challenge is that implementing the CDDF methodology is computationally taxing: our person-day sample reflects over 40 billion person-day observations.[15] We address this challenge as follows. First, we replace county, state-by-month, and month-by-year fixed effects with month, year, and division fixed effects.[16] While this requires imposing a stronger identifying assumption, we have verified that same-day mortality estimates with division fixed effects are similar to those with county fixed effects (see Online Appendix Table A12). Second, we implement the method in parallel on a dedicated server with 1 TB of memory and 32 processors. Under these conditions, the heterogeneity analysis requires four weeks to run for a single "split" of the data. Third, we do not repeat the entire analysis 100 times to account for splitting uncertainty. Instead, we estimate a single regression. This is reasonable because our sample is enormous (orders of magnitude larger than the sample sizes of the applications in CDDF) and thus is unlikely to exhibit much splitting uncertainty.

The sample used to estimate the results reported in Table 6 includes over 20 billion observations. We compute the regressions by partitioning the sample into 250 equally sized pieces and estimating 250 separate regressions. Table 6 reports the average of the 250 point estimates from those regressions. We calculate standard errors by taking the average of the 250 corresponding standard errors and then dividing by the square-root of 250. We use the same methodology to estimate equation (6). (The estimates of equation (6) are reported in Figure 5 and Online Appendix Table A14.)

---

[15] While it is possible to apply the CDDF methodology to county-level rather than person-level data, a county-level heterogeneity exercise is much less interesting because it necessitates averaging characteristics across all the beneficiaries in a county, thereby eliminating a lot of relevant variation.

[16] By design, the machine learning algorithm we employ accounts for interactions between fixed effects, so this alteration is less restrictive than it would be in a regression setting. The nontrivial change is the removal of state and county fixed effects. However, even in a regression setting this removal does not appear to matter much (see Online Appendix Table A12).

We perform the following permutation test to confirm that we are not understating the magnitude of our standard errors. For each of our 250 partitions, we randomly permute the outcome variable 100 times and estimate 100 new regressions. We then calculate an average point estimate for each permutation. This generates placebo distributions for β_1 and β_2 that are centered around zero. The estimates of β_1 and β_2 reported in column (1) of Table 6 are larger than all 100 of these placebo estimates, providing further evidence that our estimates are statistically significant.

E. Potential for bias in estimating life-years lost

We propose a new methodology that applies machine learning to estimate life-years lost and is less prone to bias than previous methods. In this Appendix, we present a framework illustrating why the traditional method of estimating life-years lost is likely to produce upwardly biased estimates. We highlight an additional assumption required to eliminate this bias and explain how our approach is more likely to meet this assumption than prior approaches. This conceptual framework motivates the data and methods we use in the main text to estimate life-years lost to fine particulate matter.

Our empirical model includes observations for individual *i* for all periods *t* up to and including the period of the individual's death. Let $c_{it} > 0$ be the expected number of remaining life-years for individual i in period t if she does not die at time *t*. For simplicity, we assume that c_{it} is unaffected by pollution.[17] Let d_{it} be an indicator equal to one if individual *i* dies in period t and equal to zero otherwise. Then $L_{it} = c_{it} d_{it}$ is the number of life-years lost due to the death of this individual in period t so that $L_{it} = c_{it}$ if the individual dies in period t, and $L_{it} = 0$ if the individual survives past period t.

We assume that exposure to PM 2.5 is assigned randomly, which effectively is the case under our instrumental variables approach. Suppose the relationship between PM 2.5 and life-years lost is governed by the following model:

$$L_{it} = \alpha + \gamma_i \text{PM2.5}_{it} + \eta_{it}. \quad \text{(A8)}$$

This model recognizes that the effect of PM 2.5 exposure on life-years lost, γ_i, can vary across individuals. The error term η_{it} represents factors other than pollution that affect mortality. By assumption of random assignment, these factors are uncorrelated with PM2.5_{it}.

Letting $\overline{\gamma} = E(\gamma_i)$, we can rewrite equation (A8) as

$$L_{it} = \alpha + \overline{\gamma} \text{PM2.5}_{it} + (\gamma_i - \overline{\gamma}) \text{PM2.5}_{it} + \eta_{it} \quad \text{(A9)}$$

[17] As in other studies, we focus on estimating the *immediate* effects of pollution exposure on life-years lost. It is also possible that exposure reduces an individual's remaining life expectancy, c_{it}, without killing her during the time window we focus on. In that case, our life-years lost estimates can be interpreted as lower bounds.

$$= \alpha + \overline{\gamma}\text{PM2.5}_{it} + v_{it},$$

where $v_{it} = (\gamma_i - \overline{\gamma})\text{PM2.5}_{it} + \eta_{it}$. Because PM 2.5 is assigned randomly, if L_{it} is perfectly observable, the standard identifying assumption $E[v_{it}|PM2.5_{it}] = 0$ holds, and OLS estimation of equation (A9) will identify $E[\hat{\gamma}] = E(\gamma_i) = \overline{\gamma}$. Thus, the presence of heterogeneous treatment effects does not, by itself, pose a problem for unbiased estimation of the average treatment effect provided that the treatment is exogenous. In practice, counterfactual life expectancy, c_{it}, is predicted rather than observed, reintroducing the potential for biased estimates of the average treatment effect, even when appropriate instruments for the treatment exist. Let $\hat{c}_{it} = \hat{g}(Z_{it})$ be the prediction of remaining life-years for individual *i* at time *t* generated by some model, $\hat{g}$, using covariates Z_{it} (e.g., age and sex). Let $u_{it} = (\hat{c}_{it} - c_{it})$ be the measurement error in this estimate so that $u_{it} > 0$ indicates the model has overestimated the true counterfactual life expectancy. We assume that the life-expectancy model is correct on average for the population (i.e., that $E(u_{it}) = 0$). The estimated number of life-years lost due to death is denoted as $\hat{L}_{it} = \hat{c}_{it}d_{it}$. Thus, $\hat{L}_{it} - L_{it} = (\hat{c}_{it} - c_{it})d_{it} = u_{it}d_{it}$. Rewriting this as $\hat{L}_{it} = L_{it} + u_{it}d_{it}$, the analog of equation (3) that the researcher can estimate with observable data is:

$$\begin{aligned} \hat{L}_{it} &= \alpha + \gamma_i\text{PM2.5}_{it} + u_{it}d_{it} + \eta_{it} \\ &= \alpha + \overline{\gamma}\text{PM2.5}_{it} + u_{it}d_{it} + v_{it} + \eta_{it}, \end{aligned} \tag{A10}$$

where u_{it} is unobservable. Unbiased estimation of $\overline{\gamma}$ in equation (A10) requires that $\text{E}(u_{it}d_{it} + v_{it} + \eta_{it}|PM2.5_{it}) = 0$. By assumption, $\text{E}(v_{it} + \eta_{it}|PM2.5_{it}) = 0$, which allows us to simplify this requirement to $\text{E}(u_{it}d_{it}|PM2.5_{it}) = \text{E}(u_{it}|PM2.5_{it})\text{E}(d_{it}|PM2.5_{it}) + Cov(u_{it}, d_{it}|PM2.5_{it}) = 0$. Since $PM2.5_{it}$ is randomly assigned, $E(u_{it}|PM2.5_{it}) = E(u_{it}) = 0$. Thus, the key requirement for unbiased estimation is that $Cov(u_{it}, d_{it}|PM2.5_{it}) = 0$. Unfortunately, this condition is unlikely to hold in most research settings. If death is a function of PM 2.5 exposure, and if unobserved health characteristics (such as latent heart disease) that make an individual more likely to die from PM 2.5 exposure also lead to overestimation of that individual's remaining lifespan (i.e., $u_{it} > 0$), then $Cov(u_{it}, d_{it}|PM2.5_{it})$ will be positive, not zero.

A natural interpretation of this issue is that it is one of selection bias: less healthy individuals are more likely to be "selected" into death due to PM 2.5 exposure, even after controlling for observables, which leads to upward bias in their estimated remaining life expectancy. However, not all unobserved health-related factors lead to bias in the estimation of $\overline{\gamma}$. It is only when these unobserved factors also make the individual more vulnerable to death via PM 2.5 exposure that the estimation error in life expectancy becomes correlated with the treatment among those who die. In other words, even if the treatment is randomly assigned, if there is heterogeneity in the propensity to die from PM 2.5 exposure, then death, and by extension life-years lost, need not be random with respect to individuals' characteristics. In this sense, the problem we encounter here is also related to bias that can arise in the case of correlated random

coefficients (Heckman and Vytlacil 1998; Wooldridge 2003), although it is more complicated in our case since there is the intervening step that the effect of the random coefficient operates by affecting who dies. This correlation may persist even when using valid instrumental variables.

To summarize, unbiased estimation of life-years lost requires modeling life expectancy using the same factors that make people susceptible to dying from pollution. We address this challenge by harnessing the comprehensive health and demographic information available in the Medicare dataset to generate relatively precise predictions of counterfactual life expectancy. In other words, we minimize the magnitude of the measurement error represented by u_{it} in equation (A10).

APPENDIX REFERENCES

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45(1): 5–32.

Chen, Tianqi and Carlos Guestrin. 2016. "XGBoost: A Scalable Tree Boosting System." arXiv: 1603.02754v3.

Chernozhukov, Victor, Mert Demirer, Esther Duflo, and Ivan Fernandez-Val. 2018. "Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments." National Bureau of Economic Research Working Paper 24678.

Einav, Liran, Amy Finkelstein, Sendhil Mullainathan, and Ziad Obermeyer. 2018. "Predictive Modeling of US Health Care Spending in Late Life." *Science* 360 (6396): 1462–1465.

Harrell, Fran, Kerry Lee, and Daniel Mark. 1996. "Tutorial in Biostatistics Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors." *Statistics in Medicine* 15 (4): 361–387.

Heckman, James and Edward Vytlacil. 1998. "Instrumental Variables Methods for the Correlated Random Coefficient Model: Estimating the Average Rate of Return to Schooling When the Return is Correlated with Schooling." *Journal of Human Resources* 33 (4): 974–987.

Ishwaran, Hemant and Udaya B. Kogalur. 2007. "Random Survival Forests for R." *R News* 7 (2): 25-31.

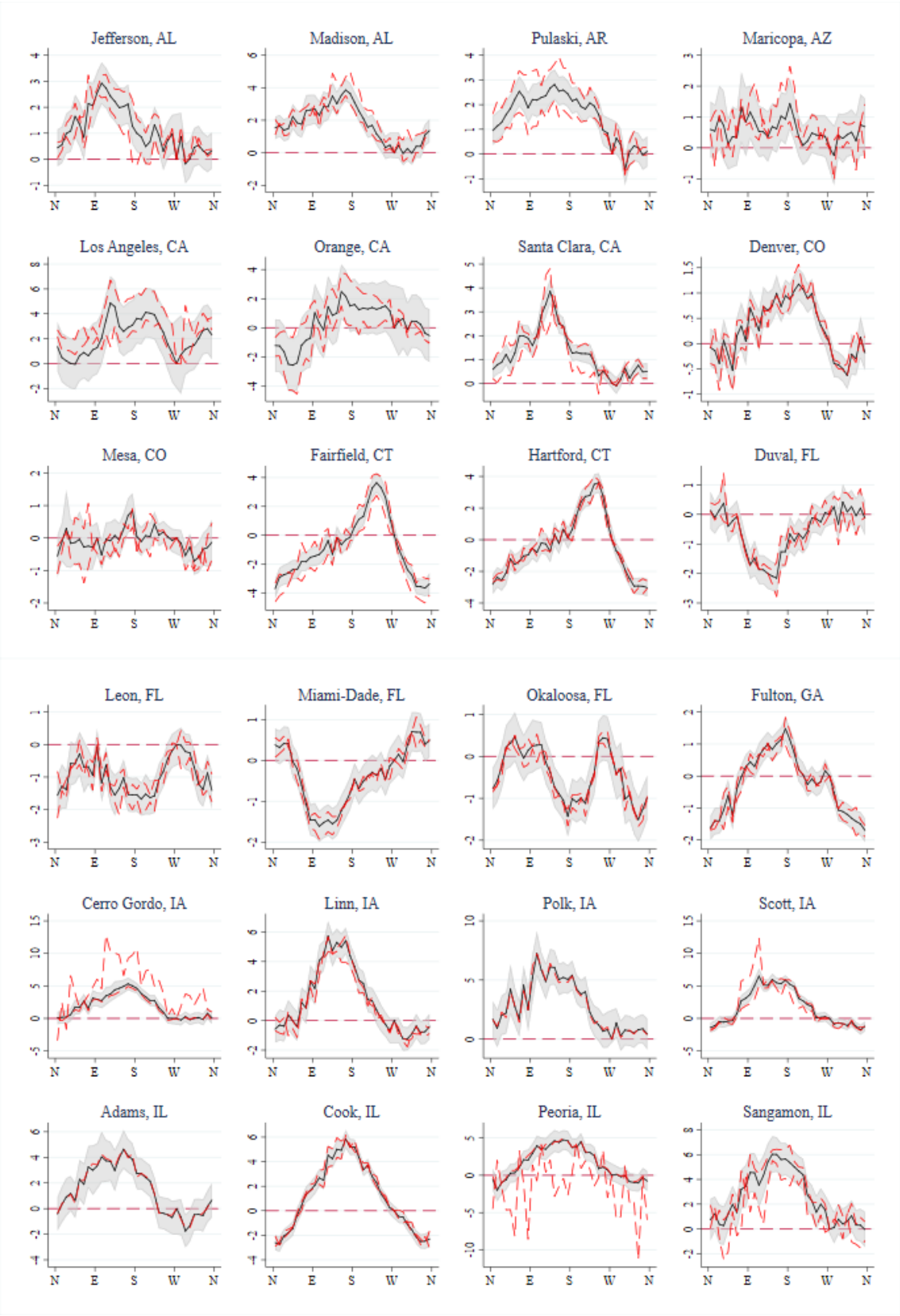
Ishwaran, Hemant, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. 2008. "Random Survival Forests." *The Annals of Applied Statistics* 2 (3): 841-860.

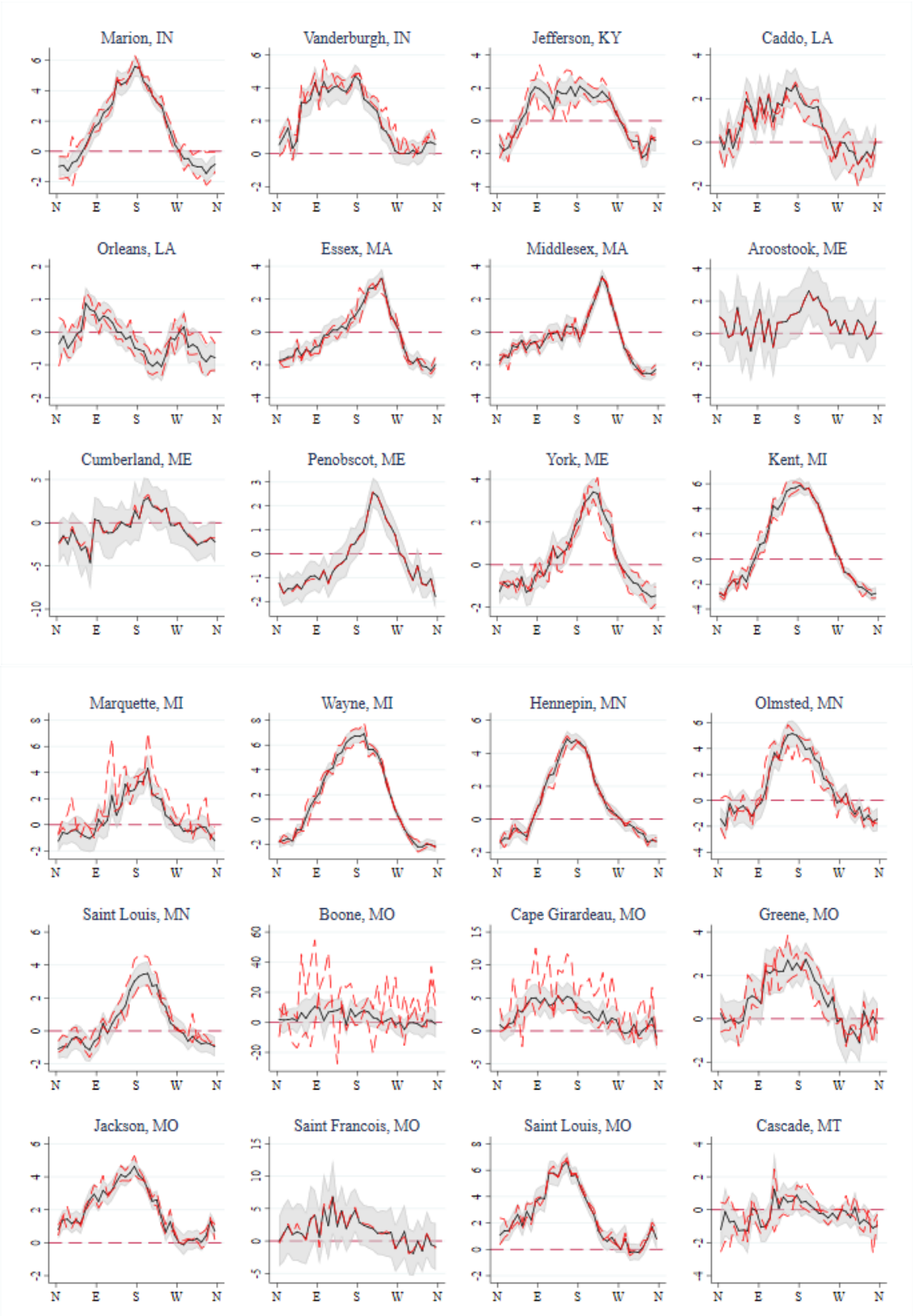
Simon, Noah, Jerome Friedman, Trevor Hastie, and Rob Tibshirani. 2011. "Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent." *Journal of Statistical Software* 39 (5): 1–13.

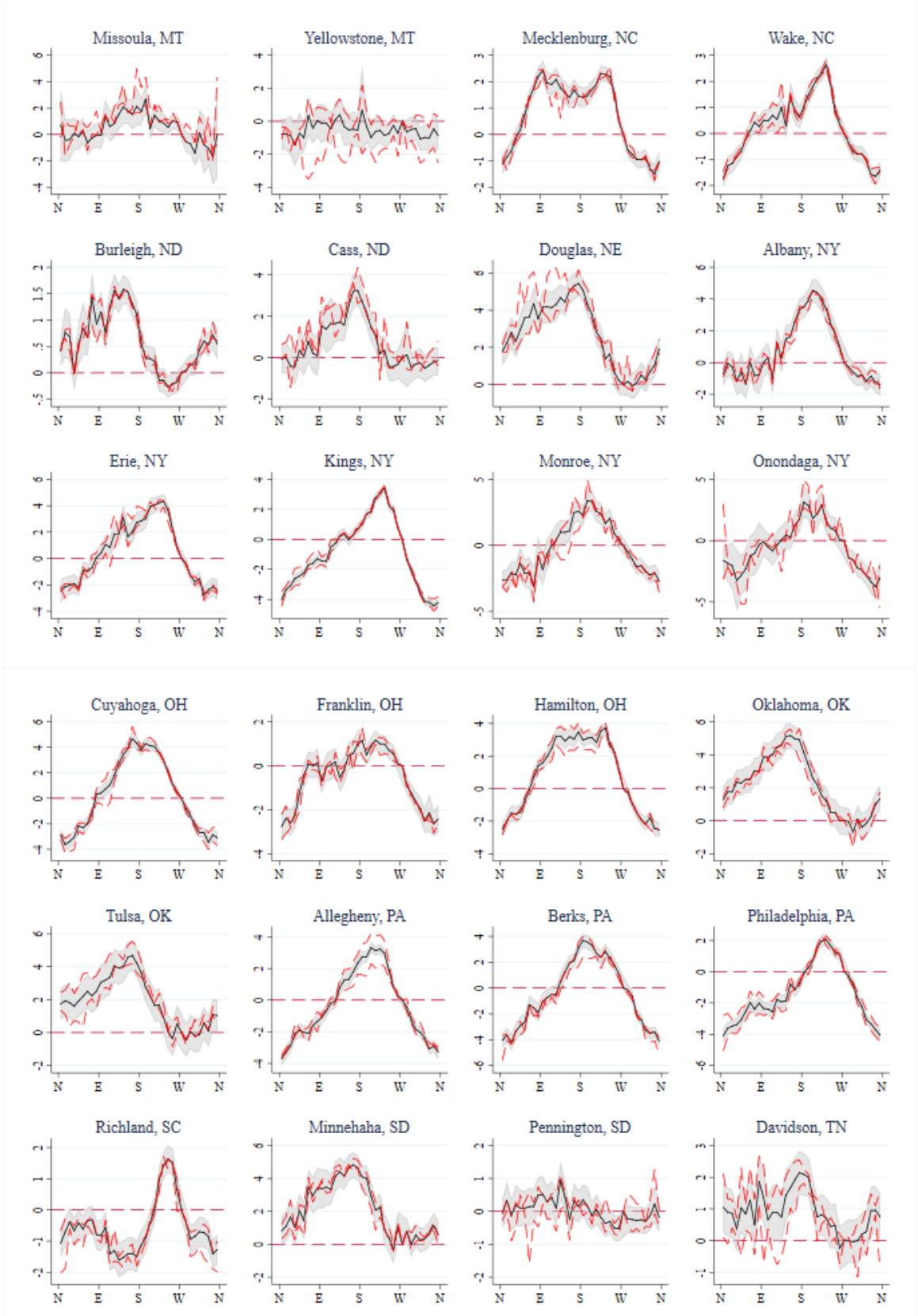
Tibshirani, Robert. 1997. "The Lasso Method for Variable Selection in the Cox Model." *Statistics in Medicine* 16 (4): 385–395.

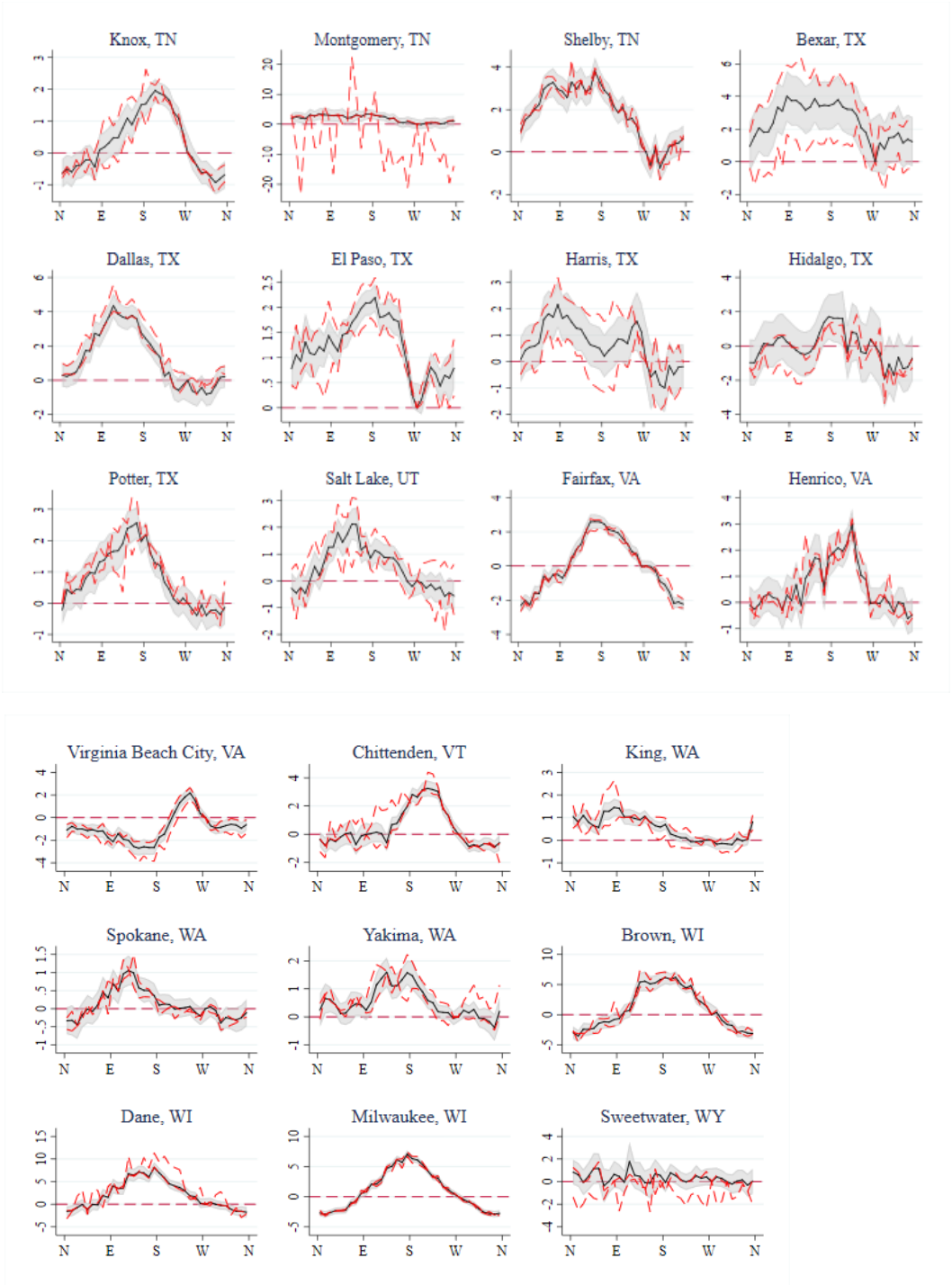
Wooldridge, Jeffrey. 2003. *Introductory Econometrics: A Modern Approach*. South-Western College Publishing.

APPENDIX FIGURES

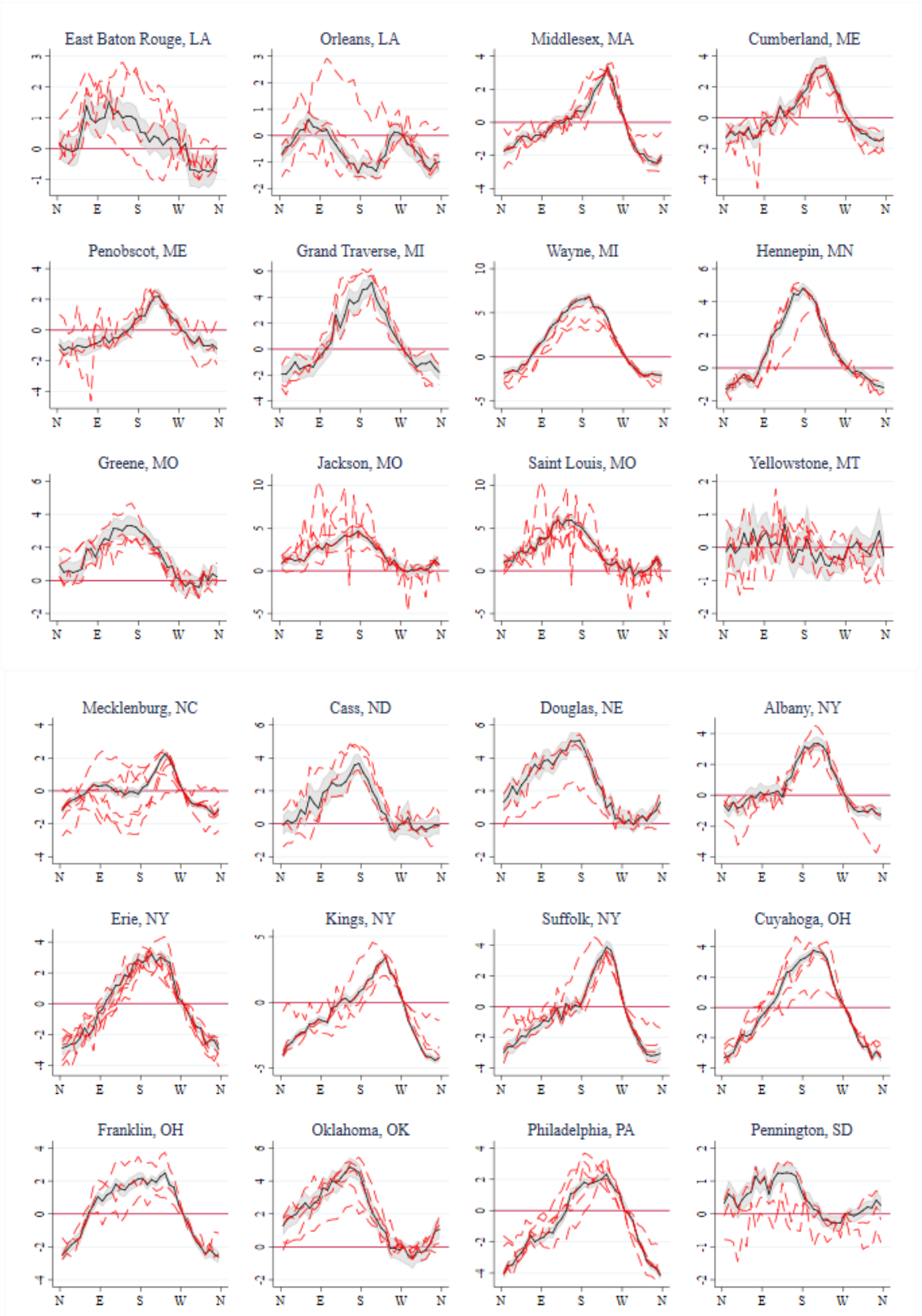


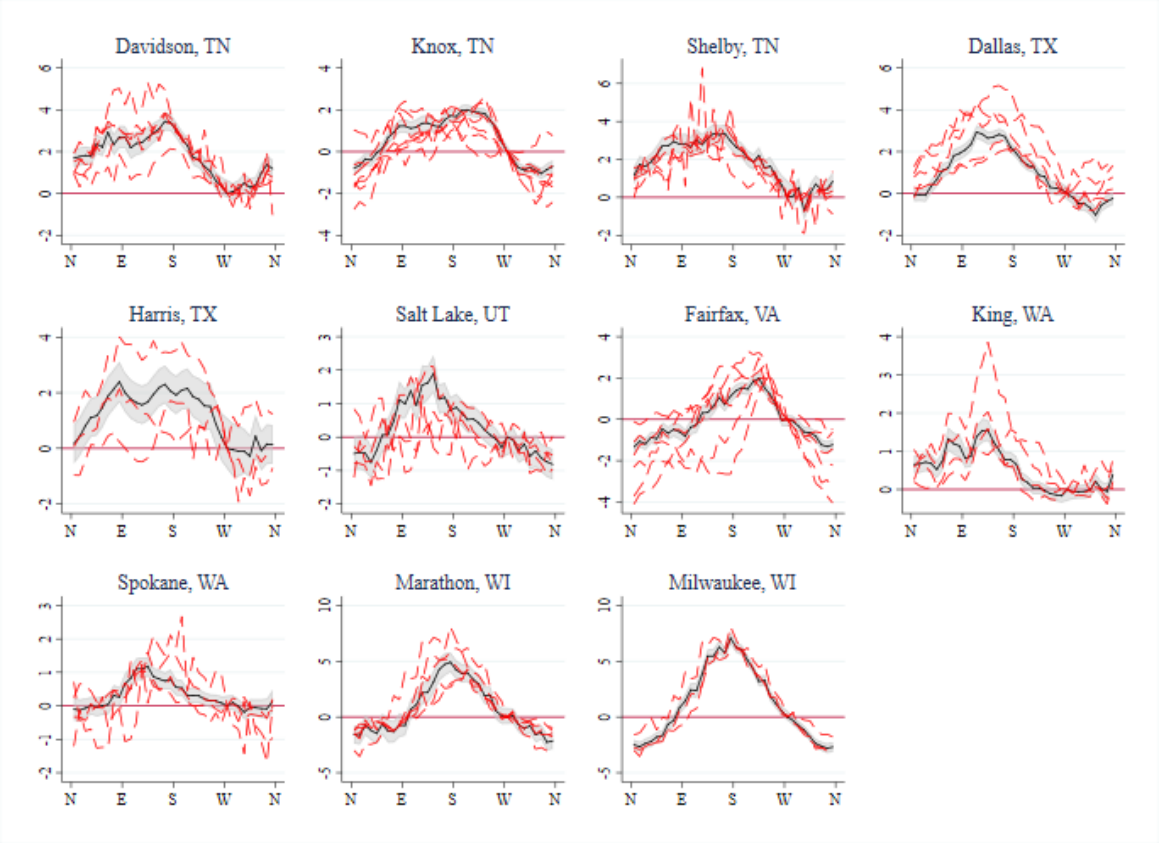

Marion, IN
Vanderburgh, IN
Jefferson, KY
Caddo, LA
Orleans, LA
Essex, MA
Middlesex, MA
Aroostook, ME
Cumberland, ME
Penobscot, ME
York, ME
Kent, MI
Marquette, MI
Wayne, MI
Hennepin, MN
Olmsted, MN
Saint Louis, MN
Boone, MO
Cape Girardeau, MO
Greene, MO
Jackson, MO
Saint Francois, MO
Saint Louis, MO
Cascade, MT
N E S W N


Missoula, MT
Yellowstone, MT
Mecklenburg, NC
Wake, NC
Burleigh, ND
Cass, ND
Douglas, NE
Albany, NY
Erie, NY
Kings, NY
Monroe, NY
Onondaga, NY
Cuyahoga, OH
Franklin, OH
Hamilton, OH
Oklahoma, OK
Tulsa, OK
Allegheny, PA
Berks, PA
Philadelphia, PA
Richland, SC
Minnehaha, SD
Pennington, SD
Davidson, TN
N E S W N

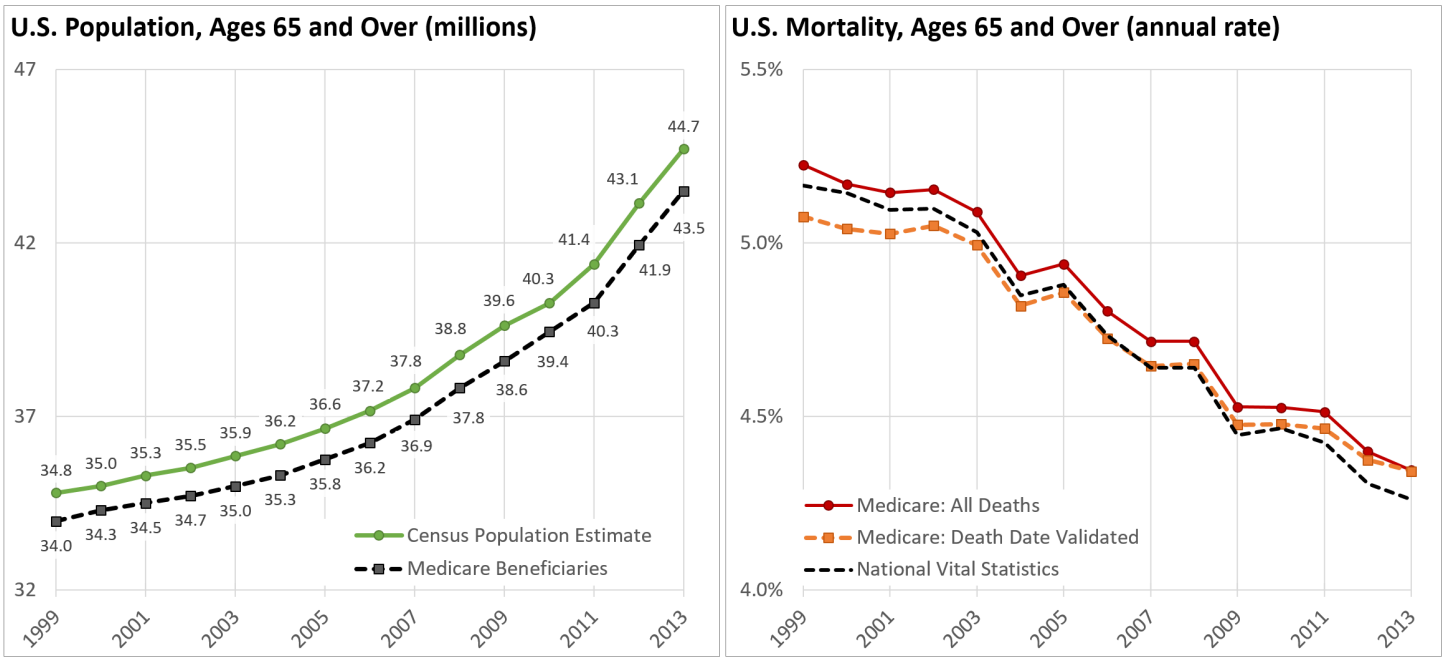


Appendix Figure A1. First-stage estimates, by monitor group. Figure plots estimates of equation (A1) for the monitor groups in the 100-monitor group classification employed in the main text. Gray area represents the 95 percent confidence interval for the overall estimate (solid black line). Dashed red lines display estimates for two subgroups to which counties in each group were randomly assigned. Graph titles report the most populous county in the group. Graphs are ordered alphabetically by state and county. Seven monitor groups with fewer than 1,000 PM 2.5 readings are not shown. Two subgroups are omitted due to insufficient number of observations (one in the Sangamon, IL group and one in the Potter, TX group).


East Baton Rouge, LA
Orleans, LA
Middlesex, MA
Cumberland, ME
Penobscot, ME
Grand Traverse, MI
Wayne, MI
Hennepin, MN
Greene, MO
Jackson, MO
Saint Louis, MO
Yellowstone, MT
Mecklenburg, NC
Cass, ND
Douglas, NE
Albany, NY
Erie, NY
Kings, NY
Suffolk, NY
Cuyahoga, OH
Franklin, OH
Oklahoma, OK
Philadelphia, PA
Pennington, SD
N E S W N



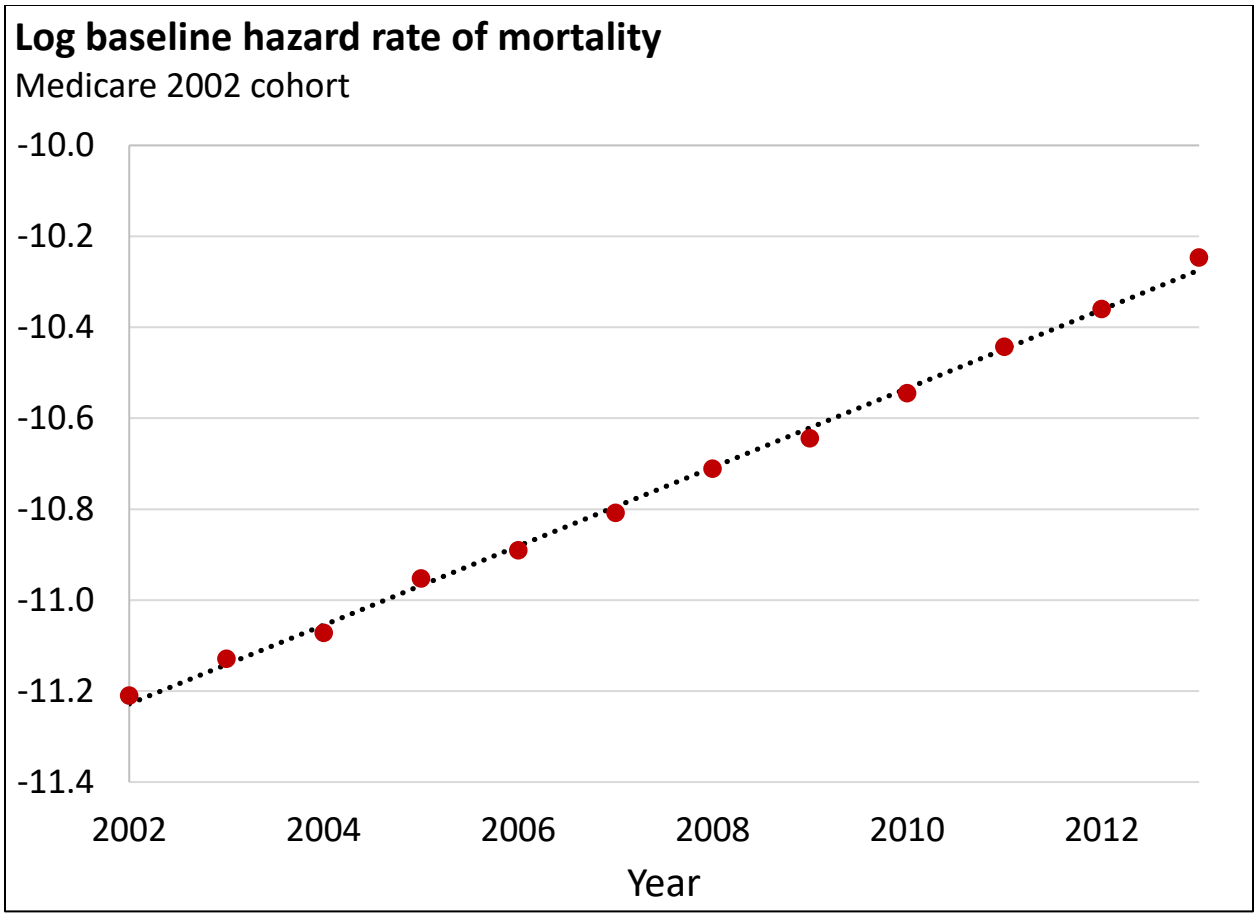
Appendix Figure A2. A comparison of first-stage estimates for the 50 and 100 monitor group specifications. Figure plots estimates of equation (A1) for each monitor group in the 50-monitor group classification and for corresponding monitor groups in the 100-monitor group classification. Gray area represents the 95 percent confidence interval for the 50-monitor-group estimate (solid black line). Dashed red lines correspond to point estimates for all monitor groups in the 100-monitor-group classification that have at least one monitor in common with the group from the 50-monitor-group classification. Graph titles report the most populous county in the group. Graphs are ordered alphabetically by state and county. Three of the 50 pollution monitor groups with fewer than 1,000 PM 2.5 readings are not shown.



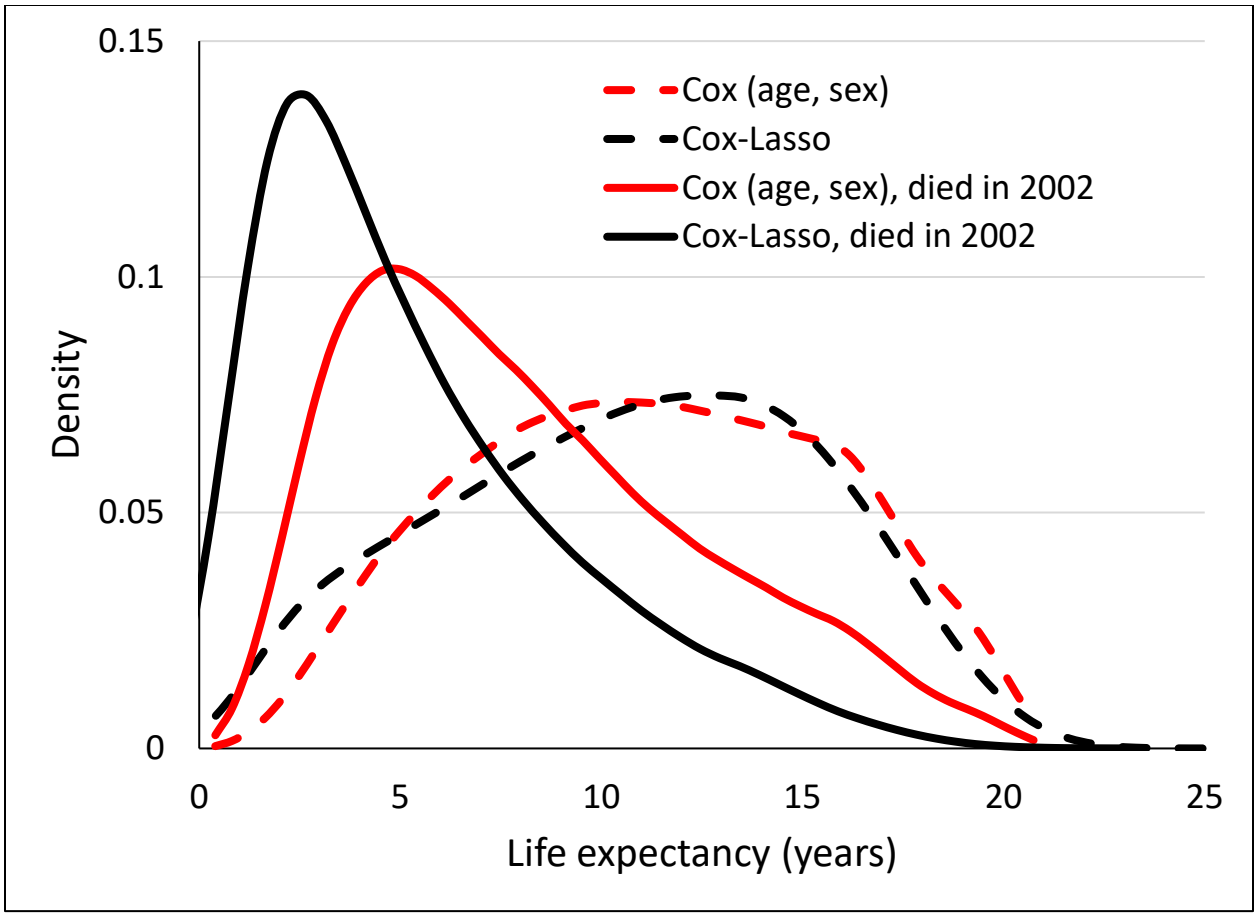
Appendix Figure A3. Population and Mortality Among US Elderly, 1999–2013.

Left Panel: Census population estimates come from the Compressed Mortality File 1999–2016 on CDC WONDER Online Database, released June 2017. These population figures are estimates of the July 1 resident population in each year except 2000 and 2010; for those two years, population figures are April 1 Census counts. Medicare beneficiaries for a given calendar year include all individuals ages 65 and over in the corresponding annual Medicare enrollment file, limited to those who turned 65 before July 1 of the year and have a ZIP code of residence located in the 50 states or the District of Columbia.

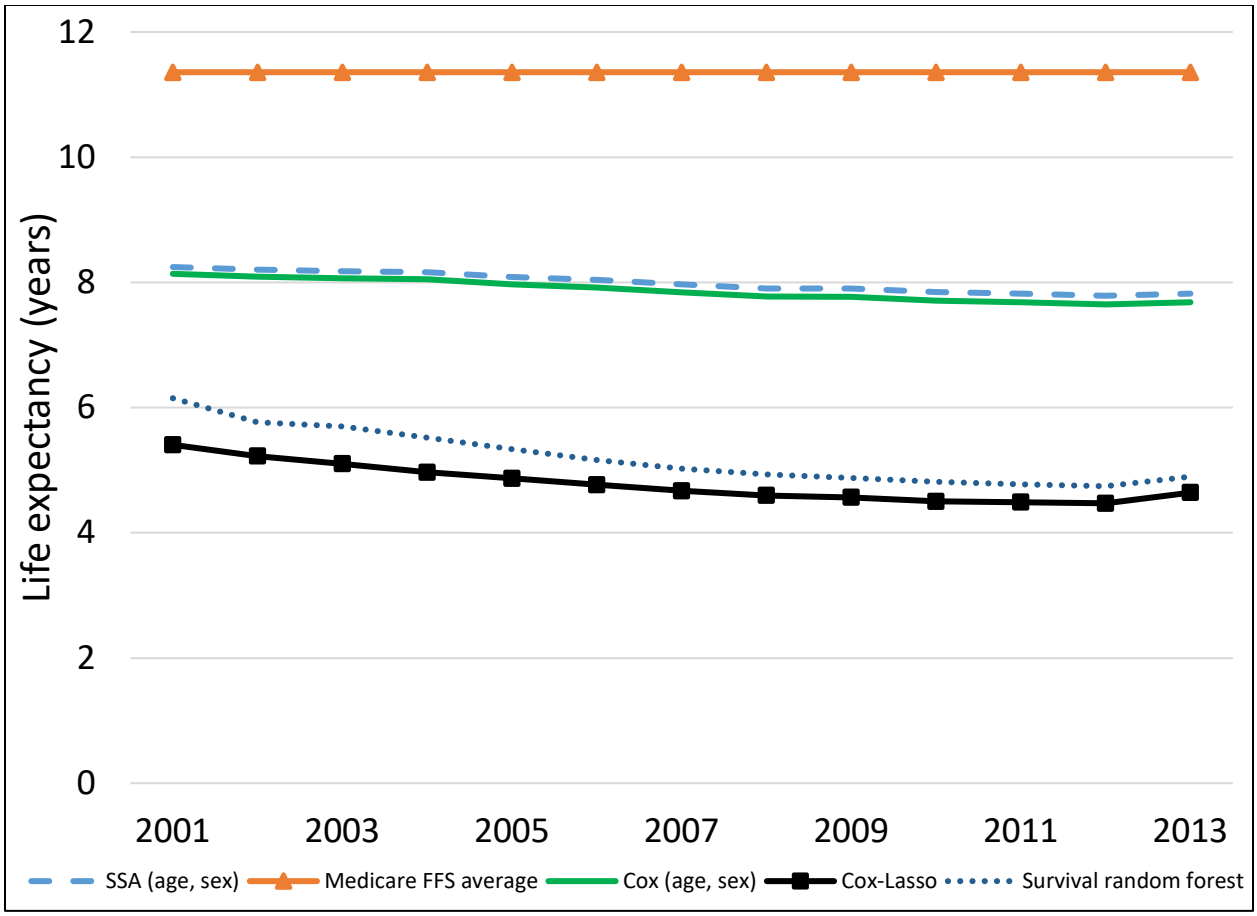
Right Panel: National Vital Statistics mortality data come from the Compressed Mortality File (CMF), which is produced by the NCHS and is based on death certificates filed in the 50 states and the District of Columbia. To obtain National Vital Statistics mortality rates, we divide total CMF deaths among the 65 and over population in a given year by the Census population estimates shown in the *Left Panel*. The dashed lines report annual mortality rates based on death dates recorded in the Medicare annual enrollment files. The figure reports both the total mortality rate in the Medicare sample ("Medicare: All Deaths"), as well as the mortality rate among the analytical sample used in the paper ("Medicare: Death Date Validated"), which excludes individuals who have a validated death that year but do not have a validated death *date* flag.



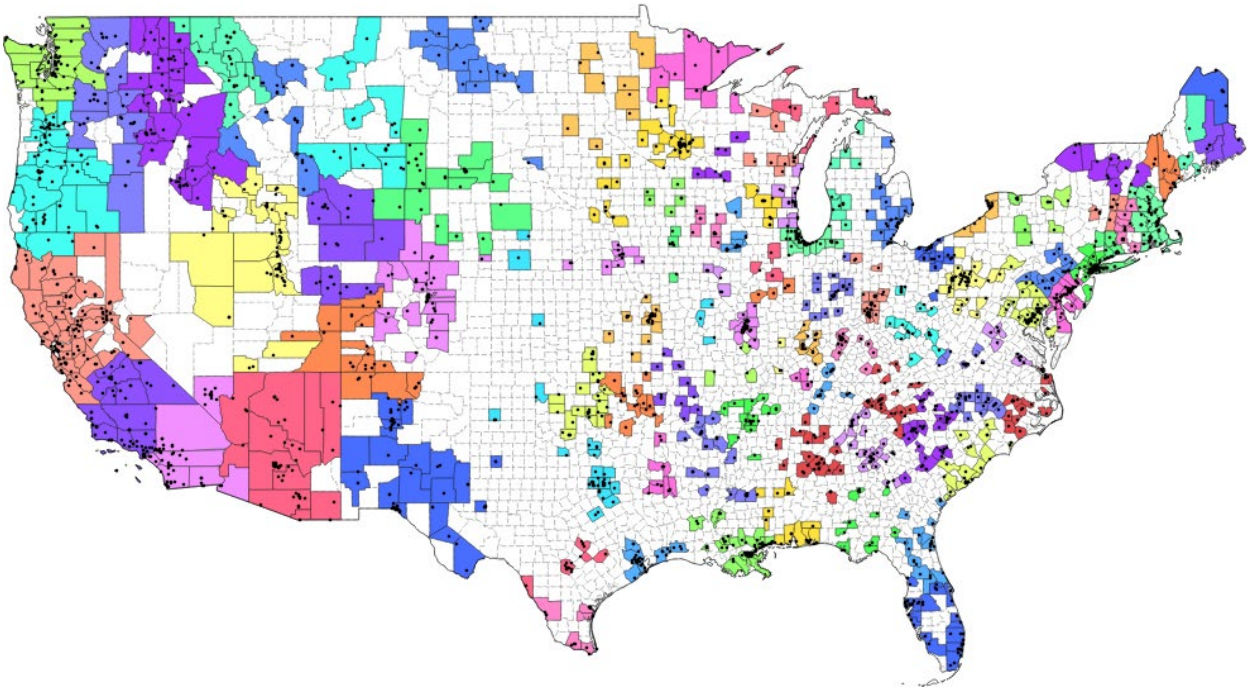
Appendix Figure A4. Log of the baseline hazard rate for the Medicare 2002 cohort. The red points in the figure correspond to the log of the baseline hazard rate of mortality for the Medicare 2002 cohort, as estimated by equation (A3) when using age and sex as controls. The coefficients on age and sex were estimated using the Cox proportional hazards model (A2). The figure also displays a dotted line of best fit.



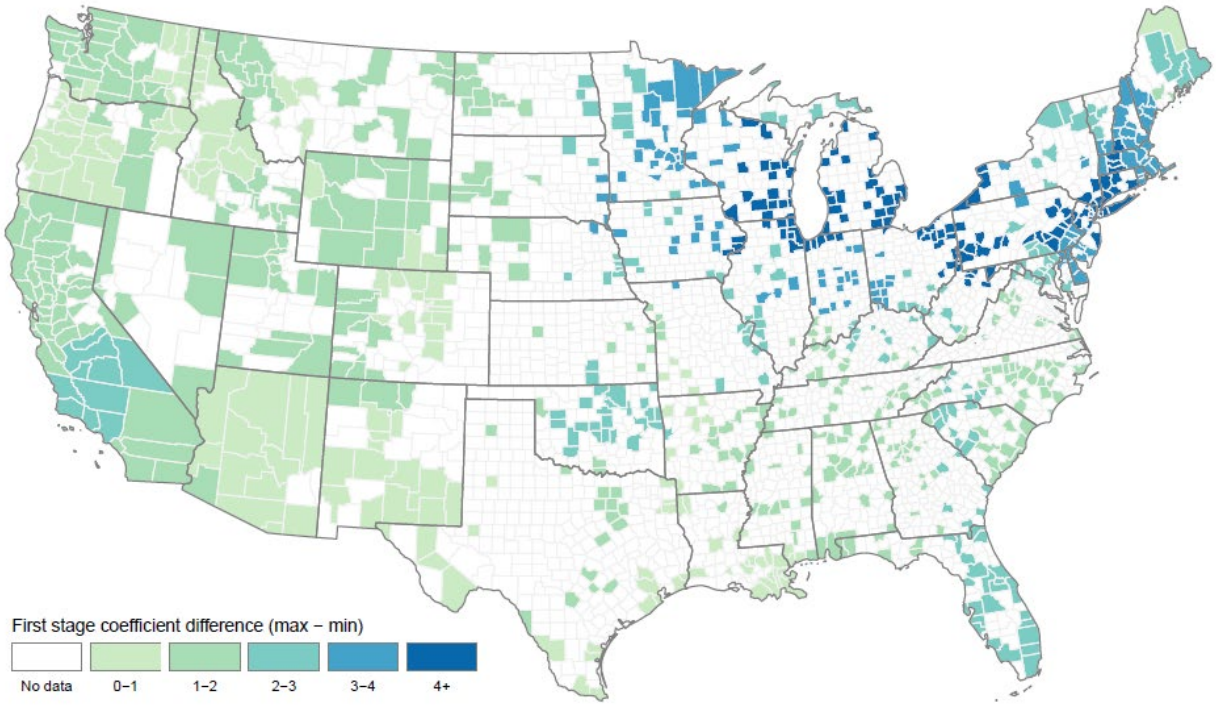
Appendix Figure A5. Kernel density plot of life expectancy estimates for Medicare beneficiaries alive on January 1, 2002. The dashed lines display the distributions of life expectancy for all Medicare beneficiaries alive on January 1, 2002. The solid lines limit the distribution to the subset of those beneficiaries who later died during the 2002 calendar year. The red lines display estimates from a Cox proportional hazards model that includes only age and sex as regressors. The black lines display estimates generated by estimating model (A3) using Cox-Lasso with 1,062 regressors.



Appendix Figure A6. Average ex ante life expectancy for Medicare fee-for-service beneficiaries who later die within one year, 2001–2013. Estimates for "Medicare FFS average" are produced by estimating equation (A2) with no covariates. Estimates for "Cox (age, sex)" are produced by estimating (A2) using only age and sex as predictors. Estimates for "Cox-Lasso" are produced by estimating (A4) with 1,062 included regressors. Estimates for "SSA (age, sex)" are obtained from the 2011 period life table for the Social Security area population (source: https://www.ssa.gov/oact/STATS/table4c6.html, accessed August 7, 2015). Estimates for "Survival random forest" are produced using the same predictors as Cox-Lasso.



Appendix Figure A7. Counties assigned to each monitor group. Different colors correspond to different monitor groups. White corresponds to counties not assigned to any monitor group due to lack of monitors. Black dots represent PM 2.5 pollution monitors.



Appendix Figure A8. Difference in First-Stage Coefficient Compliers Map. This figure displays the magnitude of the difference between the highest and the lowest coefficients among the four wind direction variables in equation (2). The darker-shaded counties have a larger difference in these coefficients, meaning that a change in wind direction has a larger effect on pollution exposure in these counties compared to the lighter-shaded counties. We refer to the darker-shaded counties as the "complier" counties.