eAppendix for: Impact of National Ambient Air Quality Standards nonattainment designations on particulate pollution and health

A Details and Tools For Constructing the Analysis Data Set

A.1 Data Sources

The tools to construct the analysis data include:

1. The AREPA R Package (available at: `https://github.com/czigler/arepa`)

2. Several R scripts (available at: `https://github.com/czigler/PM2.5-Nonattainment`)

3. Several raw data files (available at: `https://dataverse.harvard.edu/dataverse/pm97naaqs`). Note that Medicare data are not provided, rather, we provide simulated data of the same format used for linking the data sets as described below.

To assemble the analysis data set, we first obtained the locations of all ambient $\mathrm{PM}_{2.5}$ monitors in the AQS in operation between 1997 and 2012, and enumerated which were located in areas designated as nonattainment for $\mathrm{PM}_{2.5}$ in 2005. To help ensure that monitors had data spanning all seasons, monitors were only included in a given year if the annual percentage of valid measurements for that year was at least 67%. For each monitoring location, demographic information for the surrounding general population in the year 2000 was obtained by aggregating Census variables among all zip codes with centroids located within 6 miles of the monitoring location. Climate variables for each monitoring location were obtained by averaging readings taken from all ASOS monitors located within 150km of the AQS monitoring station. Medicare health outcomes for each monitoring location were obtained by aggregating mortality and hospitalization outcomes among all Medicare beneficiaries residing in

Measure	Data Source	temporal, spatial resolution	web link
Ambient $PM_{2.5}$	EPA Air Quality System	annual, monitoring site	https://www.epa.gov/aqs
Nonattainment designations	EPA Green Book	annual, county	https://www.epa.gov/green-book
Medicare Mortality and Hospitalization	Center for Medicare and Medicaid Services	annual, zip code	https://www.cms.gov/
Population Demographics	US Census and MCDC Data Archive	year 2000, zip code tabulation area	https://www.census.gov/ http://mcdc.missouri.edu/
Smoking Rates	CDC and [3, Additional file 3]	annual, county	https://www.cdc.gov/ https://goo.gl/tNbpsS
Climate	Automated Surface Observing System	annual, monitoring site	http://www.nws.noaa.gov/asos/

a zipcode with a centroid located within 6 miles of the monitoring location. Medicare beneficiaries residing within 6 miles of more than one monitoring station were uniquely assigned to the closest monitoring station, and zip codes with fewer than 15 beneficiaries were excluded. Individual-level health data are aggregated to the level of of the monitoring location to yield outcome rates (mortality rate, hospitalization rates) for all beneficiaries residing near that location.

Finally, any monitoring location with missing $PM_{2.5}$ measurements during the baseline and follow up period or with fewer than 20 Medicare beneficiaries residing in the zip codes located within 6 miles in 2012 are excluded to yield the initial analysis data set of 829 monitoring locations.

B Details of the Propensity Score "Design" Stage

Propensity scores are estimated from a logistic regression of the form:

$$log(\frac{p(A_i = 1)}{1 - p(A_i = 1)}) = X_i\gamma \tag{A.1}$$

where $A_i = 1$ denotes that the i^{th} monitoring location lies in a nonattainment area, X_i is a vector of covariates measured at the i^{th} location, including an intercept, main effects for the variables denoted in Table 1 of the main text, and two interaction terms: one between average temperature and average relative humidity and another be-eTable 1: Number of attainment and nonattainment areas in each of the four propensity score subclasses used for confounding adjustment

	Propensity Score Quartile			
	1^{st}	2^{nd}	3^{rd}	4^{th}
Attainment	94	80	40	7
Nonattainment	7	21	61	94
Total	101	101	101	101

tween average temperature and baseline ambient $\text{PM}_{2.5}$ during 2002-2004. The unknown parameter γ is estimated with maximum likelihood to construct predicted values from model (A.1), which are estimated propensity scores.

Propensity scores are estimated across the range 0 to 1, with locations in nonattainment areas tending to have higher estimated propensity scores. Importantly, the 109 locations with estimated propensity scores greater than 0.97 were exclusively nonattainment locations, and the 316 attainment locations with estimated propensity scores less than 0.019 were exclusively in attainment areas. These observations with propensity scores that do not "overlap" with the comparison group are "pruned" from the analysis data set to prevent extrapolation of causal comparisons beyond what can be supported by the observed data [2, 4, 5]. Note that the ability of the pruning strategy to ensure comparisons are confined to locations that are actually comparable is underscored by Figure 2 of the main text. Even though geography is not explicitly considered in the propensity score pruning, the resulting analysis data set is geographically more tightly clustered around the locations of the nonattianment areas to prevent, for example, outcomes in western Iowa or northern Maine from being construed as part of the "control group" for nonattainment locations that are almost certainly not comparable. After pruning, the remaining 404 monitoring locations are grouped into four groups based on quartiles of the estimated propensity score distribution. eTable 1 lists the number of attainment and nonattainment areas in each of the four subgroups.

eTable 2: Covariate and outcome comparison summary between locations retained after propensity score pruning and locations pruned due to non-overlapping propensity scores: Variables marked with ^{a} are those included in the model that estimates the propensity score, and those marked with ^{b} are those included for additional covariate adjustment in models for pollution and health outcomes. Medicare health outcomes are listed as rates per 1000 beneficiaries for mortality and per 1000 person-years for hospitalizations.

	Attainment Areas ($n = 537$)				Nonattainment Areas ($n = 292$)			
	Retained ($n = 221$)		Pruned ($n = 316$)		Retained ($n = 183$)		Pruned ($n = 109$)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Pollution Monitoring Data								
Ambient PM2.5 2002-2004 ($\mu g/m^3$)^{ab}	12.94	1.25	10.65	1.65	13.93	1.08	15.41	1.36
% Ozone nonattainment 2005^a	61.99	48.65	8.54	28	89.07	31.29	98.17	13.48
Population Demographics (Year 2000)								
log(population)^{ab}	11.82	1.41	11.14	1.66	12.33	1.38	13.03	1.26
Completely Rural Area^a	0.05	0.23	0.1	0.3	0.01	0.1	0	0
% Urban^{ab}	80.61	26.75	73.85	30.45	88.49	19.45	95.6	8.29
% Black^b	18.98	17.04	15.62	17.85	20.59	18.57	23.56	20.65
% Hispanic^b	4.51	4.45	4.28	6.67	5.63	7.76	7.97	8.69
% HS Grad.^{ab}	30.96	6.72	31.31	6.76	31.45	6.55	29.06	8.15
Median HH Inc. (\$)^a	38397.91	8646.72	36594.68	8948.49	40892.48	11702.77	45457.2	17408.51
% Poor^{ab}	14.13	5.38	14.64	6.23	13.5	6.2	13.9	7.18
% Female	51.56	1.28	51.44	1.59	51.76	1.69	51.67	1.08
% Occupied Housing^{ab}	91.88	4.47	89.8	8.52	92.43	3.21	92.41	3.18
5-Year Migration Rate^a	0.48	0.07	0.49	0.08	0.46	0.07	0.47	0.08
Median House Value (\$)^a	104426.49	45537.69	93905.49	32599.74	115520.8	53009.24	153942.61	103056.83
Smoking Rate^a	0.27	0.03	0.26	0.03	0.27	0.03	0.26	0.03
Climate (Years 2004 - 2006)								
Avg. Dew Point (°F)^a	45.79	5.26	46.73	9.37	44.82	3.27	44.32	2.94
Avg. Temperature (°F)^a	55.58	5.85	56.54	9.98	54.26	3.77	53.46	3.18
Avg. Rel. Humidity (%)^{ab}	71.66	1.85	72.76	1.7	71.87	1.7	72	1.48
Baseline Medicare Characteristics (Year 2004)								
Total Medicare Benef. 2004^{ab}	10637.62	9504.98	8373.79	10465.32	15184.84	16590.16	17014.87	17036.36
Avg. Medicare Age 2004 (years)^{ab}	75.23	0.87	75.16	0.99	75.36	0.96	75.59	1.08
% Female Medicare Benef. 2004^{ab}	59.57	2.56	58.75	3.09	59.81	2.35	59.81	2.17
% White Medicare Benef. 2004^{ab}	85.98	15.2	87.93	15.56	83.04	21.03	78.77	21.96
% Black Medicare Benef. 2004^{ab}	12.09	14.84	9.83	14.39	14.49	20.4	16.56	20.21
Mortality^a	53.79	9.37	51.78	8.95	53.21	6.35	53.65	9.21
All CVD^a	112.53	27.36	105.65	25.3	119.38	24.14	122.32	29.25
Respiratory^a	33.27	13.58	30.78	12.31	34.01	10.67	35.38	14.22
COPD	11.16	7.4	10.14	5.91	11.03	4.92	11.66	5.9
CV Stroke	19.73	5.61	18.16	6.02	20.56	5.22	20.09	4.69
Heart Failure	25.51	8.75	22.5	9.2	27.7	9.3	30.01	11.78
HRD	15.24	5.11	14.96	4.76	16.28	3.99	15.76	3.47
Ischemic Heart Disease	30.04	11.44	29.57	9.41	30.92	9.19	30.91	8.86
Peripheral Vascular Disease	6.9	2.95	6.63	3.29	6.98	2.12	7.77	2.87
Respiratory Tract Infection	22.1	8.02	20.64	8.13	22.99	6.84	23.72	10.28
Pollution and Health Outcomes								
Ambient PM2.5 2010-2012 ($\mu g/m^3$)	10.01	1.29	8.89	1.74	10.86	1.3	11.53	1.32
Mortality 2012	49.22	6.26	47.21	8.17	47.96	6.11	48.99	13.41
COPD 2012	11.36	7.24	9.37	5.75	11.39	4.88	11.54	5.94
CV Stroke 2012	15.44	4.16	14.13	5.21	15.89	4.05	15.93	4.6
Heart Failure 2012	17.89	5.85	15.46	6.56	19.11	7.56	19.18	6.6
HRD 2012	13.96	3.94	13.48	4.86	15.36	4.23	14.26	4.09
Ischemic Heart Disease 2012	14.9	5.26	16.02	6.26	15.74	4.91	15.59	5.88
Peripheral Vascular Disease 2012	4.1	1.62	4.27	2.97	4.69	2.4	4.91	2.16
Respiratory Tract Infection 2012	16.27	6.7	14.52	6.63	16.41	5.35	15.73	5.94

C Details of the Models Used in the "Analysis Stage"

C.1 Spatial Hierarchical Model for Air Pollution

Estimates of effects of the nonattainment designations on average ambient $\mathrm{PM}_{2.5}$ in 2010-2012 are derived from a model fit to (log-transformed) potential pollution values of the form:

$$Y_i(A) = X_i\beta^A + W_i^A + \epsilon_i^A, \tag{A.2}$$

where $Y_i(A)$ denotes the potential pollution concentration that would be observed under designation A, with $A = 0$ denoting "attainment" and $A = 1$ denoting "nonattainment." In this model, X_i include dummy variables indicating propensity score subclass membership as well as main effects for the covariates that showed some sign of imbalance in the design, as denoted in Section B and in Table 1 of the main text. ϵ_i^A are normally-distributed non-spatial "nugget" errors, $\epsilon_i^A \sim N(0, \psi^A)$. W_i^A are spatial random effects defined with a multivariate gaussian process having an isotropic and stationary exponential covariance function. The spatial decay of covariances between two locations governed by parameters, θ^A, and the variances of these spatial random effects are denoted τ^A. Specifying separate models for the values of A implies the assumption that, conditional on X_i, the potential pollution concentrations under $A = 0$ and $A = 1$ are conditionally independent. Details of this model specification, including the structure of a sensitivity analysis to this conditional independence assumption, can be found in [6]. Further details of the spatial random effect specification can be found in [1].

C.2 Log-linear model for Medicare Health Outcomes

Estimates of the effects of the nonattainment designations on Medicare health outcomes derive from augmenting the spatial pollution model in (A.2) with log-linear models for the mortality and hospitalization rates. We assume conditional independence of potential health outcomes, conditional on covariates and air pollution. We also assume that potential health outcomes under a nonattainment designation status are independent of what pollution would have been under an attainment designation, after conditioning on covariates and pollution under a nonattainment designation. As a result of these assumptions, models for potential health outcomes can be written as:

$$\log(E[H_i(A)]) = \alpha_0^A + X_i\alpha_1^A + Y_i(A)\alpha_2^A + \log(N_i), \tag{A.3}$$

where $H_i(A)$ is the potential number of outcomes (mortalities or hospitalizations) at location i under designation A and $N(s)$ is the total number of Medicare beneficiaries (managed care and fee-for-service) for the mortality outcome, or person-years among fee-for-service beneficiaries for the hospitalization outcomes. X_i is a vector of covariates including dummy variables for propensity score subclass membership and main effects for the same covariates included in (A.2) and denoted in Appendix B and in Table 1 of the main text.

C.3 Missing Data

Our analysis is confronted with two types of missing ambient pollution data. Among all locations in the analysis data set, 131 (40 in nonattainment areas) had missing baseline $\mathrm{PM}_{2.5}$ measurements in 2002-2004. Missing values of average annual ambient pollution in 2002-2004 were imputed using a single posterior mean prediction from a spatial hierarchical random effects model of the form (A.2) including all baseline demographic and climate variables listed in Table 1 of the main text, as well as indicators for a nonattainment designation for $\mathrm{PM}_{2.5}$, PM_{10}, ozone, and sulfur dioxide in 2005. Among the locations retained after propensity score pruning, 67 (27 in nonattainment) used this imputed value of ambient $\mathrm{PM}_{2.5}$ in 2002-2004.

The second type of missing pollution data are missing average annual ambient concentrations during the follow up period of 2010-2012. In the entire set of analysis locations, 263 (74 nonattainment) had missing follow up pollution. Among those retained after propensity score pruning, 118 (52 in nonattainment) had missing follow up pollution. These follow up pollution measures are used as outcomes in our analysis, and are multiply imputed from the models described above as a byproduct of our Bayesian estimation procedure. No part of the analysis used locations that had both missing baseline and follow up pollution. All other covariates and outcomes listed in Table 1 were fully observed.

C.4 Bayesian Estimation and Prior Specification

Causal effects estimates are derived from the posterior distribution of the parameters in the above models, estimated with Markov chain Monte Carlo (MCMC) data augmentation algorithm that iteratively samples missing potential outcomes conditional on observed data and parameters, then samples parameters and calculates causal effect estimates conditional on "complete" data. For example, one iteration of the MCMC algorithm calculates the average difference between $Y(A=1)$ and $Y(A=0)$ in the sample of retained nonattainment locations to estimate the sample average causal effect of the nonattainment designation on ambient pollution among nonattainment areas. For a given location, at most one of these values is observed and at least one of these values is simulated form its posterior-predictive distribution. Health effects are estimated analogously. Further details appear in [6]. For the analyses of the main text, MCMC chains were run for 50,000 iterations, with the first 5,000 discarded as "burn in" and every 10^{th} posterior sample retained for posterior inference.

C.4.1 Prior distributions and outline of MCMC strategy

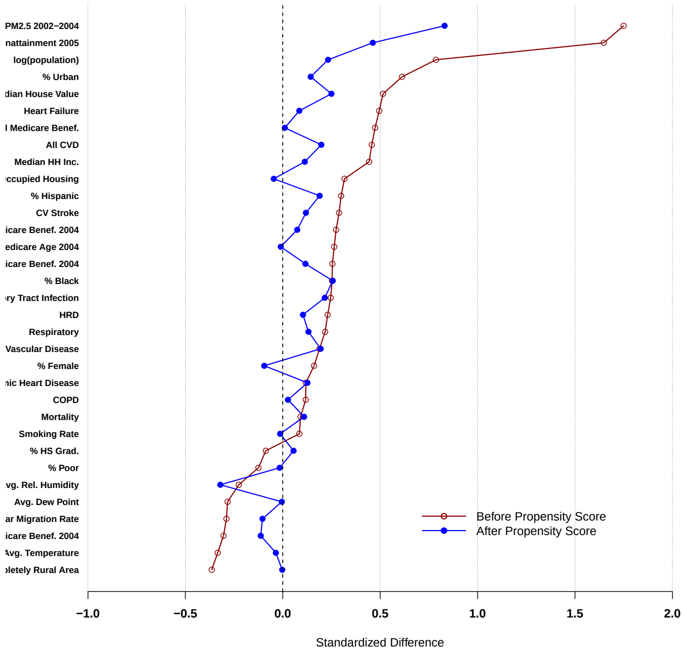
We treat the parameters $\beta^0, \beta^1, \psi^A, \psi^1, \tau^A, \tau^1, \theta^0, \theta^1, \alpha^0$ and α^1, as *a priori* independent. We specify flat priors for β, α^0 and α^1. For the ψ^A and τ^A, we specify independent inverse-gamma distributions with shape parameters set to 2 and scale parameters set to 0.5. For θ^A, we specify uniform prior distributions on the interval $(0.48, 3.61)$. Parameters for the prior distributions of ψ^A and θ^A are meant to reflect diffuse prior information within the range of plausible parameter values.

D Sensitivity Analysis without Pruning

D.1 Pollution Only Model

A sensitivity analysis analogous to that described in the main text was carried out on the entire analysis data set of 829 monitoring locations (i.e., without the propensity score pruning). This analysis uses the same propensity score estimates, but no observation is pruned due to non-overlapping propensity scores. Locations were grouped into four propensity score subclasses according to whether the estimated propensity score was: 1) ≤ 0.25; 2) > 0.25 and ≤ 0.5; 3) > 0.5 and ≤ 0.75; or 4) > 0.75. The number of attainment (nonattainment) locations in each subclass was 467 (19), 37 (17), 21 (32), 12 (224), respectively. Figure **??** shows the covariate balance when all 829 locations are grouped into a propensity score subclass, indicating substantial imbalances relative to the pruned analysis of the main text and pronounced potential for confounding. Models of the form (A.2) and (A.3) were fit, adjusting for propensity score subclass and main effects for the following covariates: $\text{PM}_{2.5}$ in 2002-2004, ozone nonattainment designation in 2005, log of population, % urban population,% black, % high school graduates, % occupied housing, average temperature, average relative humidity, and total number of, average age of, % female, % white, and % black Medicare beneficiaries in 2004, all of which had average ASMD across all subclasses of greater than 0.25 or a difference greater than 0.5 in any subclass.

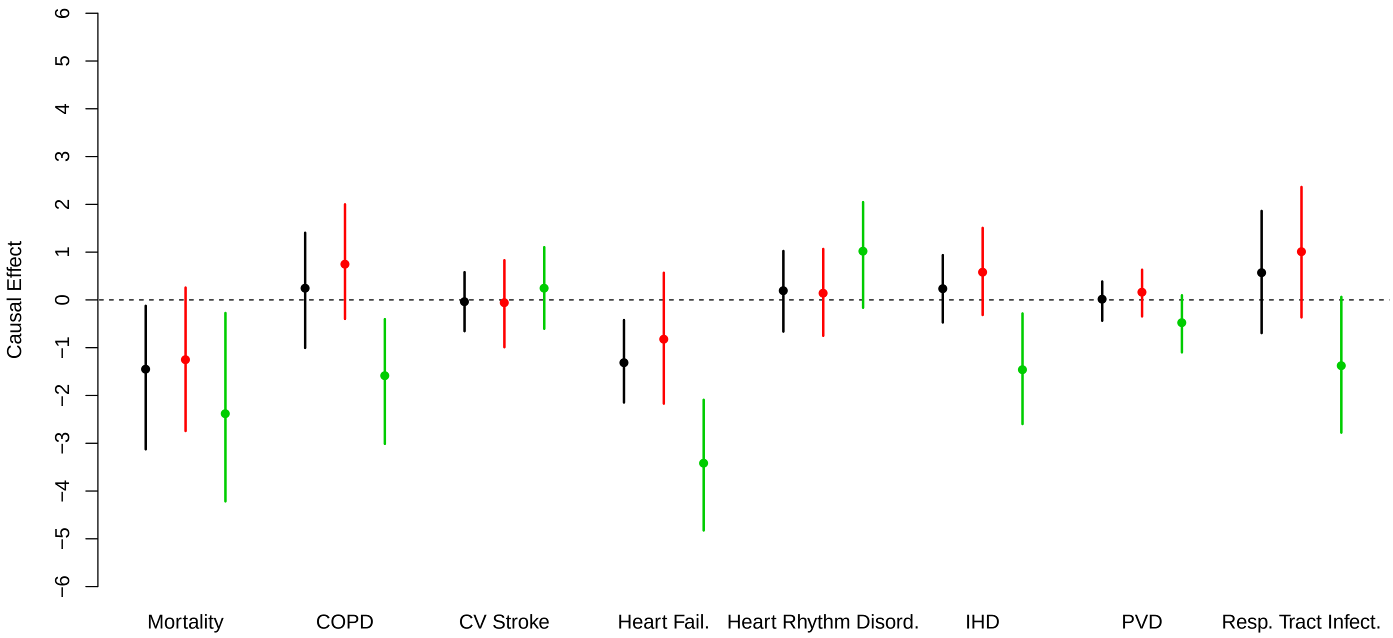
eFigure 1: Plots of average standardized mean difference in each measured covariate between attainment and nonattainment locations for the sensitivity analysis without propensity score pruning. Red line denotes differences before propensity score subclassification, blue line denotes differences within propensity score subclass.



For the unpruned analysis, the estimated causal effect of the nonattainment designations on ambient $\text{PM}_{2.5}$ in 2010-2012 among the 292 nonattainment areas was -0.459 $\mu g/m^3$, with 95% posterior interval (-1.342, 0.289). dFigure 2 summarizes posterior distributions for the average causal effects of the nonattainment desingations on the Medicare mortality rate (per 1000 beneficiaries) and hospitalization rates (per 1000 person-years) among all nonattainment locations, as well as average dissociative and associative effects. This sensitivity analysis indicates significant overall reductions in rates of all-cause mortality and hospitalizations for heart failure. Estimates of dissociative effects are all very similar to estimates of overall effects. Associative effects are more pronounced than dissociative effects for rates of all-cause mortality and COPD, Heart Failure, IHD, PVD, and RTI hospitalizations, with significantly negative estimates for all-cause mortality, COPD, Heart Failure, IHD, and RTI.

Note that the MCMC chain for the sensitivity analysis of pollution outcomes was run for 19500 iterations (after discarding 5000 as burn in) of which every 10^{th} sample was retained for posterior inference. Sensitivity analyses of health outcomes were based on MCMC chains run for between 8000 and 13000 iterations (after discarding 5000 as burn in), depending on the outcome, with every 10^{th} iteration retained for posterior inference.

eFigure 2: Posterior mean point estimates and 95% posterior probability intervals for overall, associative, and dissociative effects in the sensitivity analysis of $PM_{2.5}$ nonattainment designations without propensity score pruning.



References

[1] Banerjee, Sudipto and Gelfand, Alan E. and Finley, Andrew O. and Sang, Huiyanauthor (collaboration) (2008), "Gaussian predictive process models for large spatial data sets," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70, 825–848.

[2] Crump, Richard K. and Hotz, V. Joseph and Imbens, Guido W. and Mitnik, Oscar A.— (2009), "Dealing with limited overlap in estimation of average treatment effects," *Biometrika*, 96, 187 –199.

[3] Dwyer-Lindgren, Laura and Mokdad, Ali H. and Srebotnjak, Tanja and Flaxman, Abraham D. and Hansen, Gillian M. and Murray, Christopher JL— (2014), "Cigarette smoking prevalence in US counties: 1996-2012," *Population Health Metrics*, 12, 5.

[4] Ho, Daniel E. and Imai, Kosuke and King, Gary and Stuart, Elizabeth A.— (2007), "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference," *Political Analysis*, 15, 199 –236.

[5] King, Gary and Zeng, Langche— (2006), "The Dangers of Extreme Counterfactuals," *Political Analysis*, 14, 131 –159.

[6] Zigler, Corwin M and Dominici, Francesca and Wang, Yun— (2012), "Estimating causal effects of air quality regulations using principal stratification for spatially correlated multivariate intermediate outcomes," *Biostatistics (Oxford, England)*, 13, 289–302.