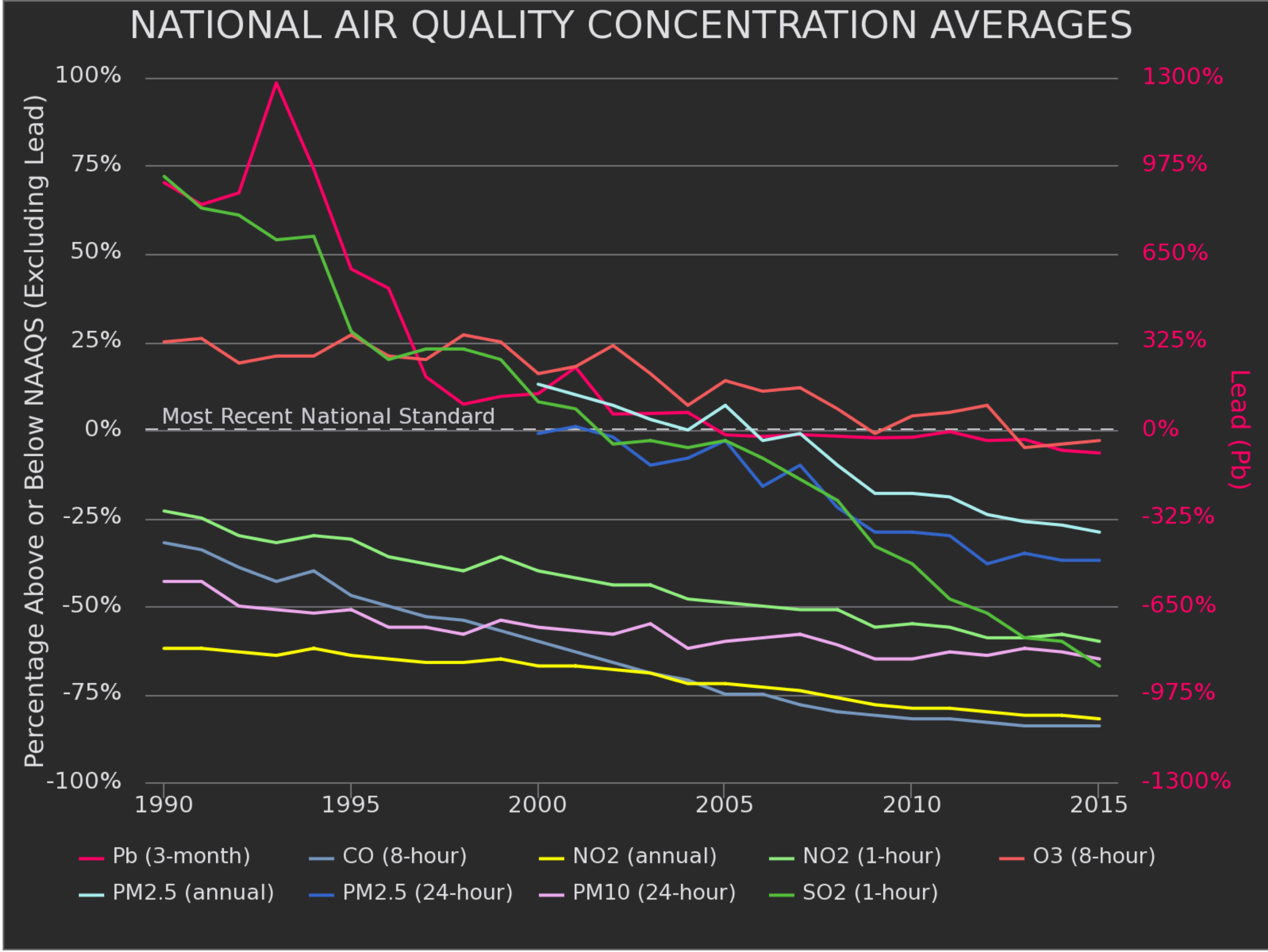
Data Science and the Environment

Francesca Dominici, PhD
Professor of Biostatistics
Harvard TH Chan School of Public Health
Co-Director Harvard Data Science Initiative

HDSI | Harvard Data Science Initiative

President Nixon is signing the Clean Air Act in 1970



NATIONAL AIR QUALITY CONCENTRATION AVERAGES
Percentage Above or Below NAAQS (Excluding Lead)
100%
75%
50%
25%
0%
-25%
-50%
-75%
-100%
Most Recent National Standard
1300%
975%
650%
325%
0%
-325%
-650%
-975%
-1300%
Lead (Pb)
1990
1995
2000
2005
2010
2015
Pb (3-month)
CO (8-hour)
NO2 (annual)
NO2 (1-hour)
O3 (8-hour)
PM2.5 (annual)
PM2.5 (24-hour)
PM10 (24-hour)
SO2 (1-hour)


PM2.5
Nickel
Mercury
CO2
NOx
SO2
...

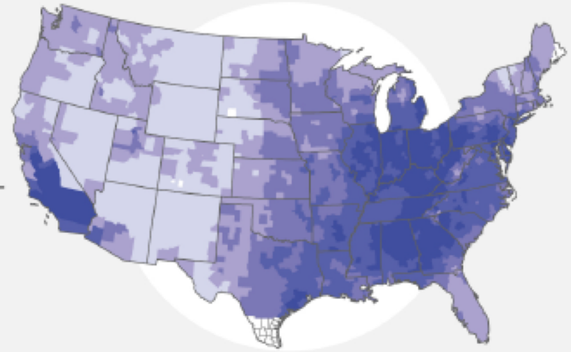
Scientific Questions

1. Is exposure to $PM_{2.5}$ below the NAAQS (35 mu_g/m^3 for short term and 12 mu_g/m^3 for long term) associated with an increase mortality risks?

2. Are some populations at higher risk than others?

RESEARCH DATA PLATFORM

EXPOSURES AND INTERVENTIONS (E OR I)



$PM_{2.5}$ exposure levels by county (average 2000-2012)

DATA SOURCES

Criteria air pollutants
EPA AQS daily average of $PM_{2.5}$, ozone, NO_2, 1995-2015;
Daily 1km x 1km predictions of $PM_{2.5}$, ozone, NO_2, 2000-2014

Methane
1km x 1km predictions at 3-day intervals, 2009-present

Weather
NOAA daily estimates (temperature, precipitation, humidity, ...) on a 0.3° grid

Power plants
EPA AMPD daily emissions, 1995-2015

Coal mines
MSHA location and producting pits, 1970-2015

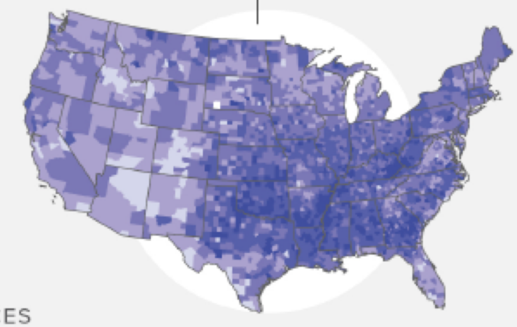
Fracking wells and disposal wells
Drillinginfo database with well location and depth, daily production

Traffic
Annual traffic counts and density from the Department of Transportation

Residential community green space
NASA vegetation index on a 250m² grid

Factrories and industrial sites
Geocoded locations of businesses

HEALTH OUTCOMES (Y)



Medicare mortality rate by county (average 2000-2012)

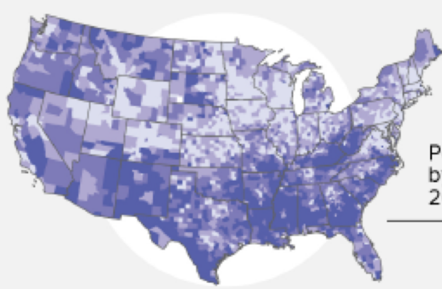
DATA SOURCES

Medicare
28 million per year, 1999-2015

Medicaid
28 million per year, low income, 2010-2011

Aetna
40 million, all ages, above-average income, 2008-2016

CONFOUNDERS (X)



Poverty prevalence by county (average 2000 and 2010)

DATA SOURCES

Individual demographics
Age, sex, race, ZIP code of residence

Individual medical history
Previous diagnoses, medications prescribed

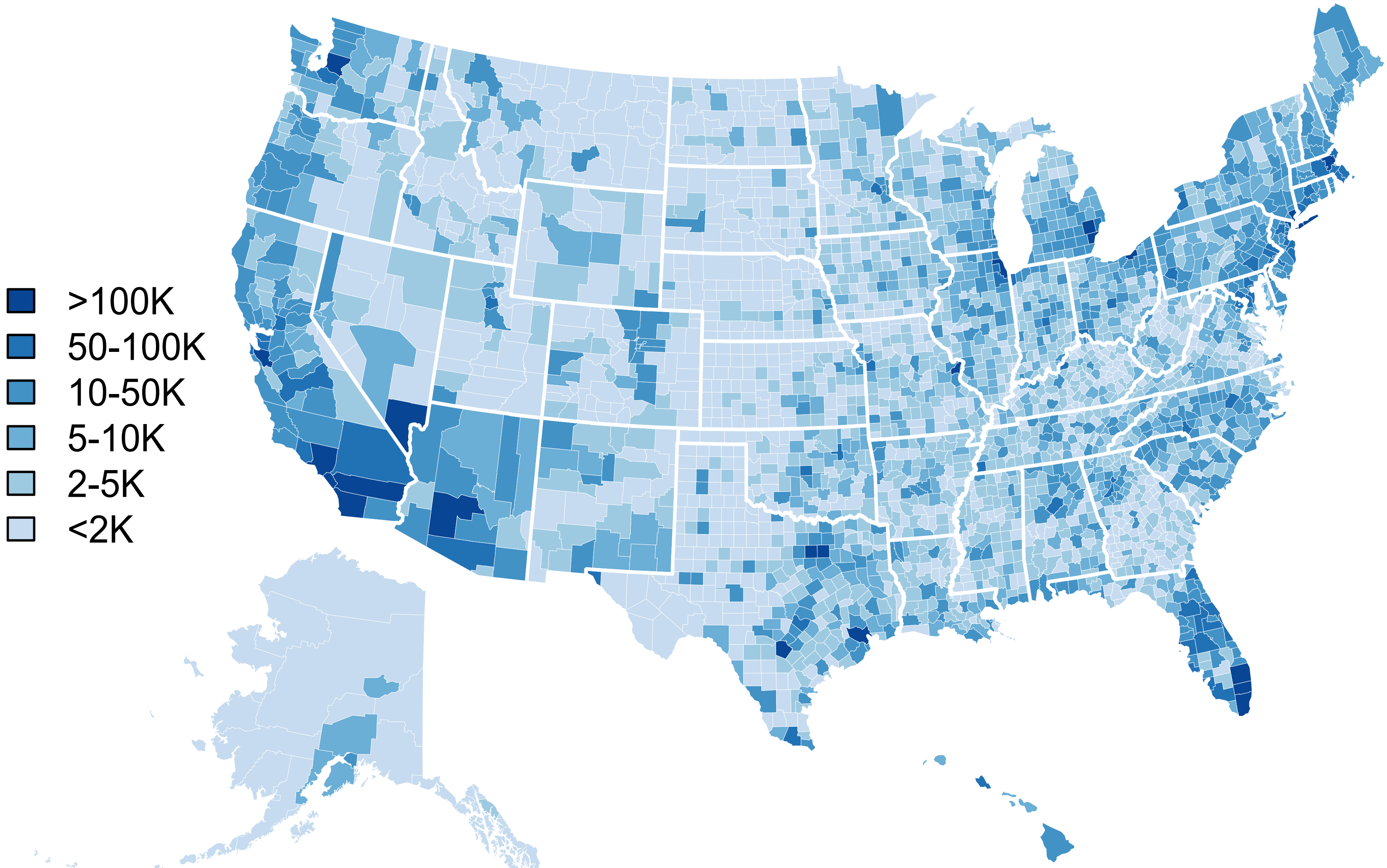
ZIP code level variables
Income, education, demographics, employment, household size

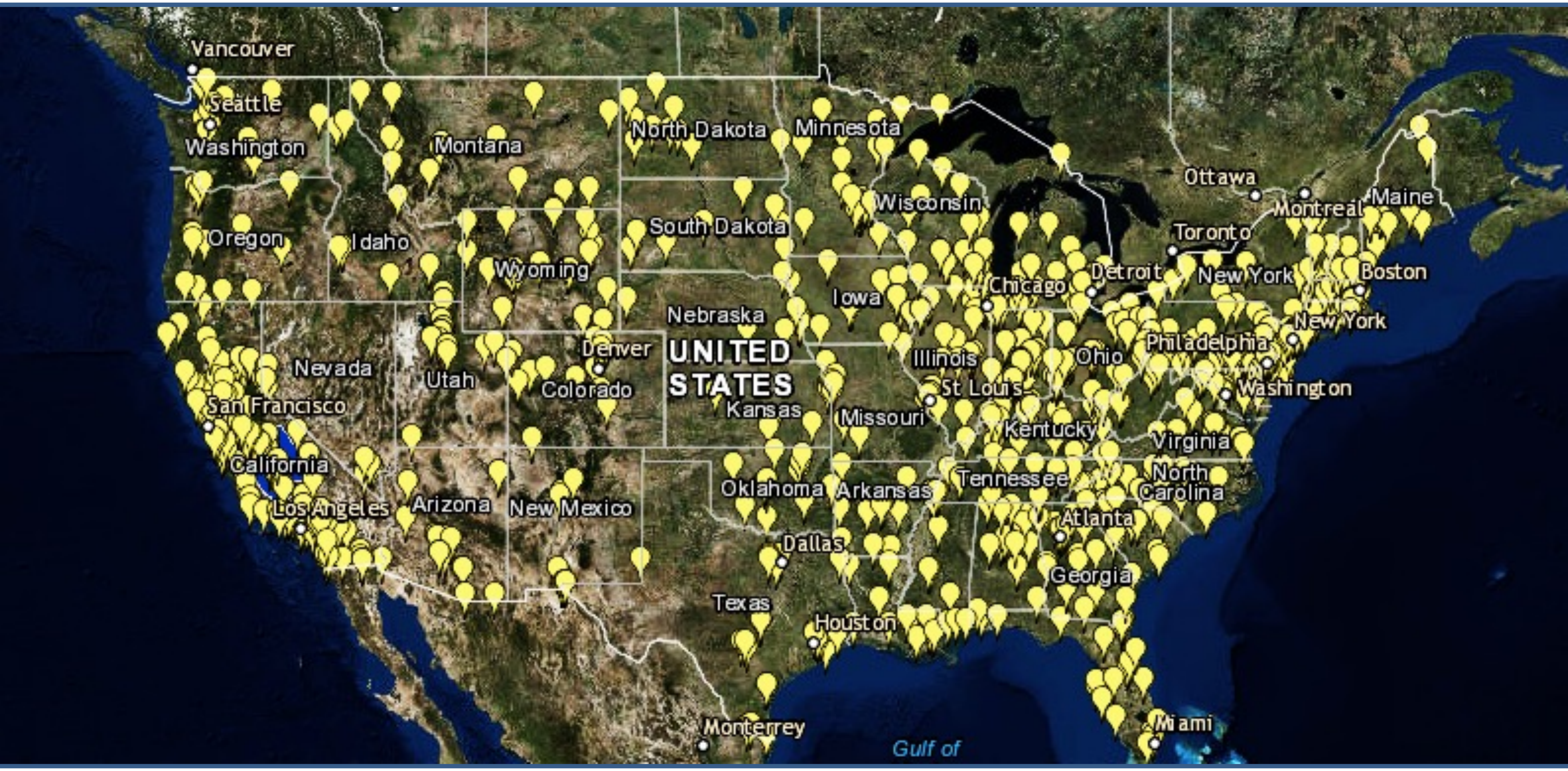
County-level variables
Crime, smoking, BMI

DATA

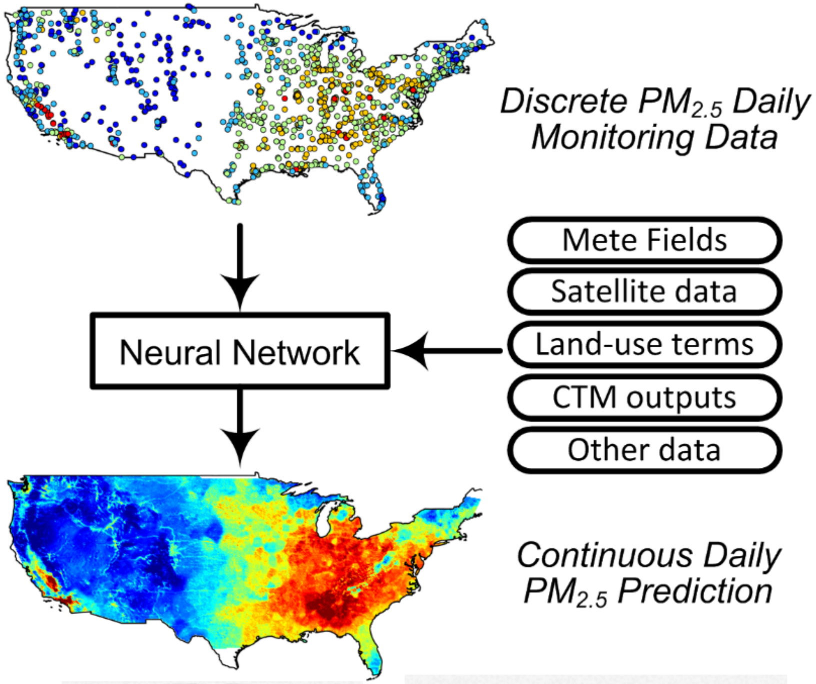
- All Medicare participants (n=67,682,479) in the continental United States from 2000 to 2012 (updating the data to 2015)
- Outcomes: all-cause mortality and cause specific hospitalization
- Individual level information: date of death, age of entry, year of entry, sex, race, whether eligible for Medicaid (proxy for SES)
- Zip code of residence and other covariates

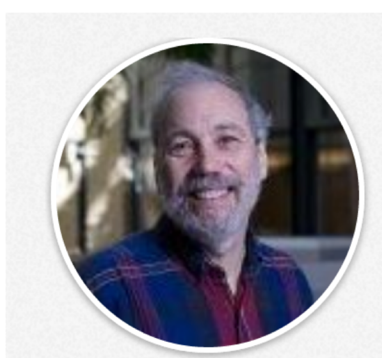
Medicare Data (open cohort of 60 million enrollees at year from 1999 to 2015)
480 person years




Vancouver
Seattle
Washington
Montana
North Dakota
Minnesota
Oregon
Idaho
Wyoming
South Dakota
Wisconsin
Ottawa
Montreal
Maine
Toronto
Boston
Detroit
New York
Chicago
Iowa
Nebraska
New York
Nevada
Utah
Denver
UNITED STATES
Illinois
Ohio
Philadelphia
Colorado
St Louis
Washington
San Francisco
Kansas
Missouri
Kentucky
Virginia
California
Tennessee
North Carolina
Oklahoma
Arkansas
Los Angeles
Arizona
New Mexico
Atlanta
Dallas
Georgia
Texas
Houston
Monterrey
Miami
Gulf of

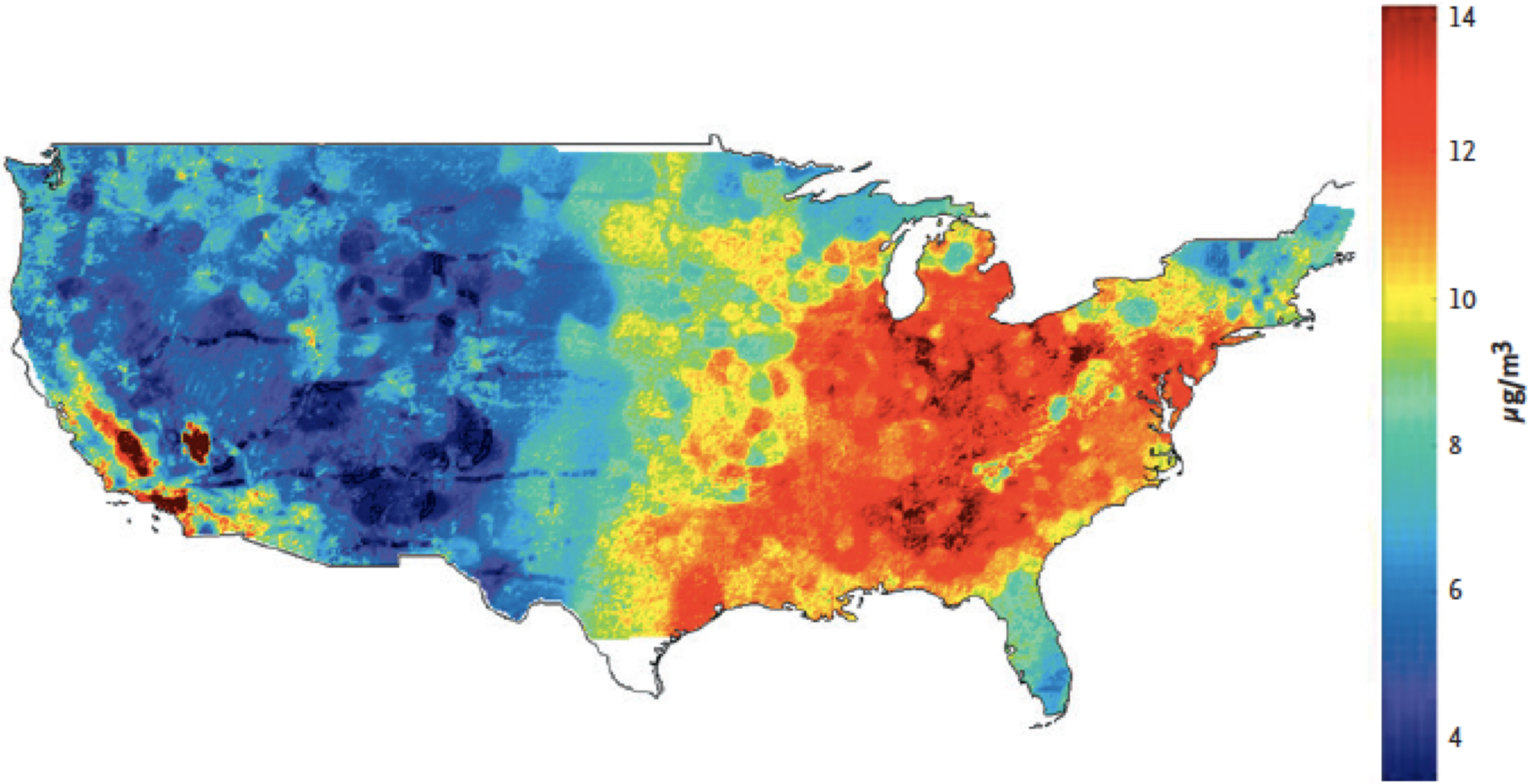
Neural Network for Exposure Prediction





- A neural network to incorporate satellite-based measurements, simulation outputs from a chemical transport model (CTM), land-use terms and other ancillary data to model monitored $PM_{2.5}$ and ozone
- Model training at monitors:
  - $PM_{2.5} \sim \beta_1 Mete + \beta_2 Satellite + \beta_3 land + \beta_4 CTM + \beta_5 others$
- Modeling prediction without monitors:
  - $\widehat{PM_{2.5}} \sim \beta_1 Mete + \beta_2 Satellite + \beta_3 land + \beta_4 CTM + \beta_5 others$

A Average Concentrations of $PM_{2.5}$



The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 JUNE 29, 2017 VOL. 376 NO. 26

Air Pollution and Mortality in the Medicare Population

Qian Di, M.S., Yan Wang, M.S., Antonella Zanobetti, Ph.D., Yun Wang, Ph.D., Petros Koutrakis, Ph.D.,
Christine Choirat, Ph.D., Francesca Dominici, Ph.D., and Joel D. Schwartz, Ph.D.

Table 1. Cohort Characteristics and Ecologic and Meteorologic Variables.

Characteristic or Variable	Entire Cohort	Ozone Concentration		$PM_{2.5}$ Concentration	
		≥50 ppb*	<50 ppb	≥12 $\mu g/m^3$	<12 $\mu g/m^3$
Population					
Persons (no.)	60,925,443	14,405,094	46,520,349	28,145,493	32,779,950
Deaths (no.)	22,567,924	5,097,796	17,470,128	10,659,036	11,908,888
Total person-yr†	460,310,521	106,478,685	353,831,836	212,628,154	247,682,367
Median yr of follow-up	7	7	7	7	7
Average air-pollutant concentrations‡					
Ozone (ppb)	46.3	52.8	44.4	48.0	45.3
$PM_{2.5}$ ($\mu g/m^3$)	11.0	10.9	11.0	13.3	9.6
Individual covariates‡					
Male sex (%)	44.0	44.3	43.8	43.1	44.7
Race or ethnic group (%)§					
White	85.4	86.6	85.1	82.0	88.4
Black	8.7	7.2	9.2	12.0	5.9
Asian	1.8	1.8	1.8	2.1	1.6
Hispanic	1.9	2.0	1.9	1.9	1.9
Native American	0.3	0.6	0.3	0.1	0.6
Eligible for Medicaid (%)	16.5	15.3	16.8	17.8	15.3
Average age at study entry (yr)	70.1	69.7	70.2	70.1	70.0

Table 2. Risk of Death Associated with an Increase of 10 μg per Cubic Meter in $PM_{2.5}$ or an Increase of 10 ppb in Ozone Concentration.*

Model	$PM_{2.5}$	Ozone
	hazard ratio (95% CI)	
Two-pollutant analysis		
Main analysis	1.073 (1.071–1.075)	1.011 (1.010–1.012)
Low-exposure analysis	1.136 (1.131–1.141)	1.010 (1.009–1.011)
Analysis based on data from nearest monitoring site (nearest-monitor analysis)†	1.061 (1.059–1.063)	1.001 (1.000–1.002)
Single-pollutant analysis‡	1.084 (1.081–1.086)	1.023 (1.022–1.024)

Increases of 10 μg/m3 in $PM_{2.5}$ and of 10 ppb in ozone were associated with increases in all-cause mortality of 7.3% (95% confidence interval [CI], 7.1 to 7.5) and 1.1% (95% CI, 1.0 to 1.2), respectively.

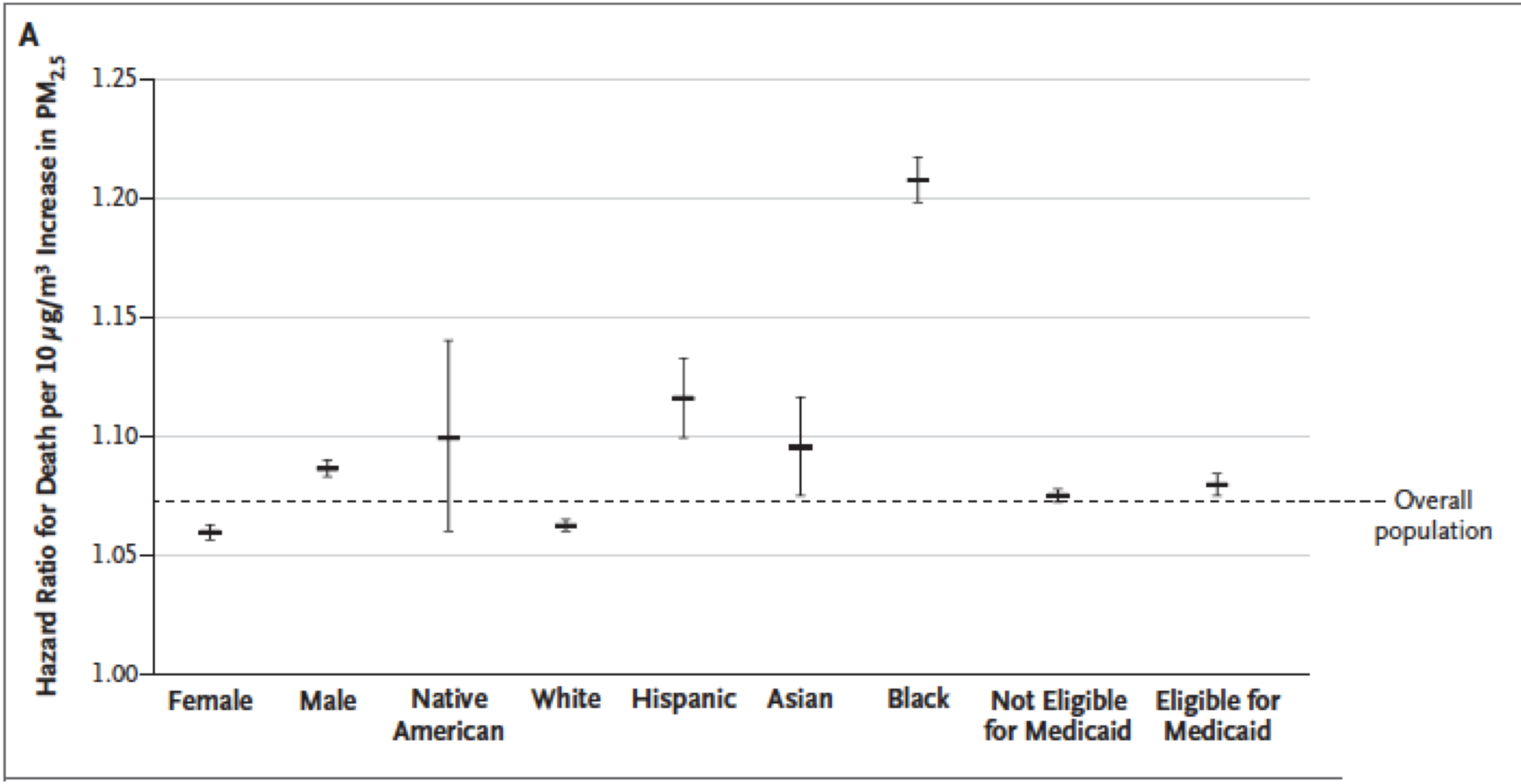
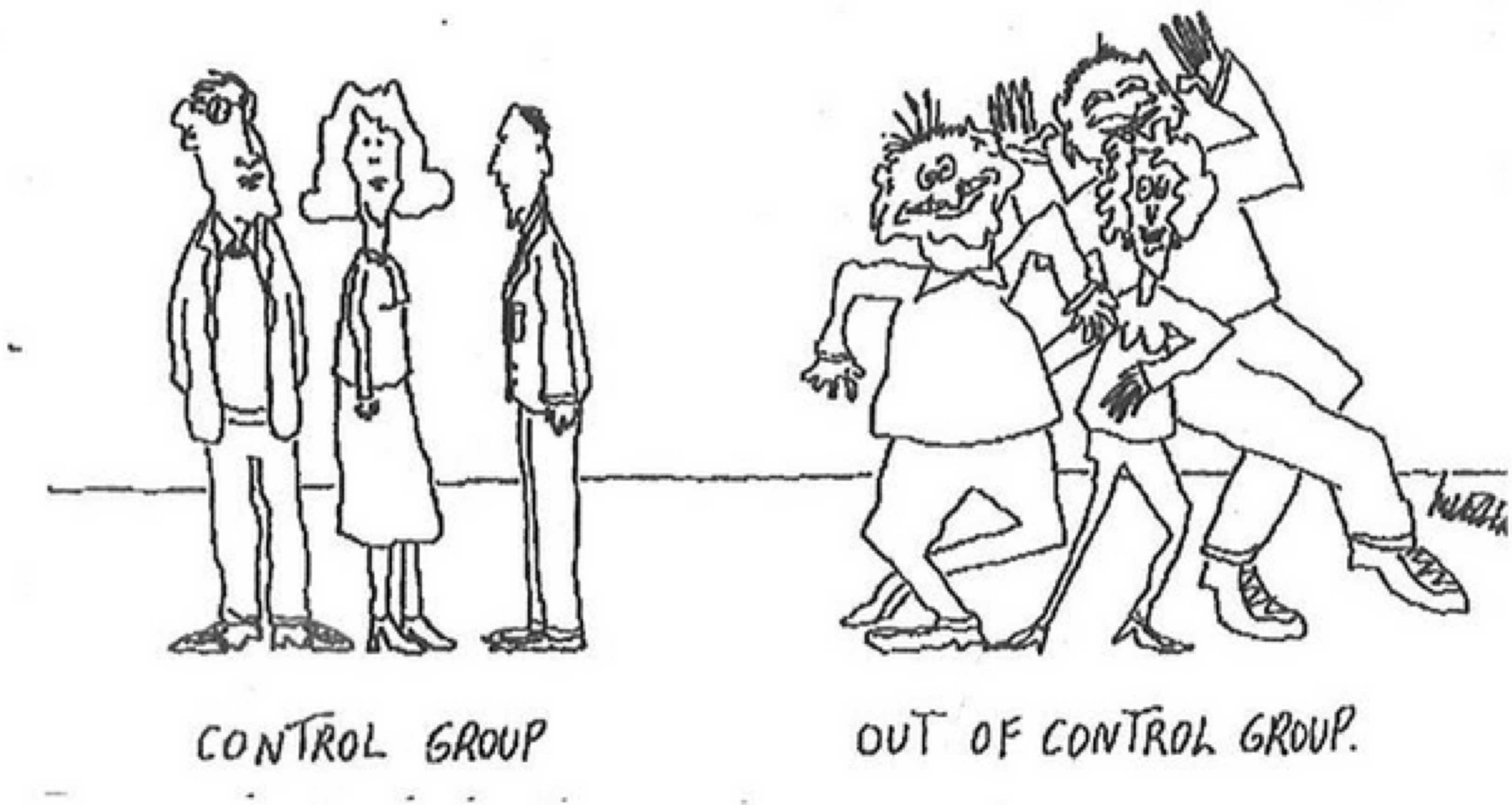


Figure 2. **Risk of Death Associated with an Increase of 10 μg per Cubic Meter in $PM_{2.5}$ Concentrations and an Increase of 10 ppb in Ozone Exposure, According to Study Subgroups.**

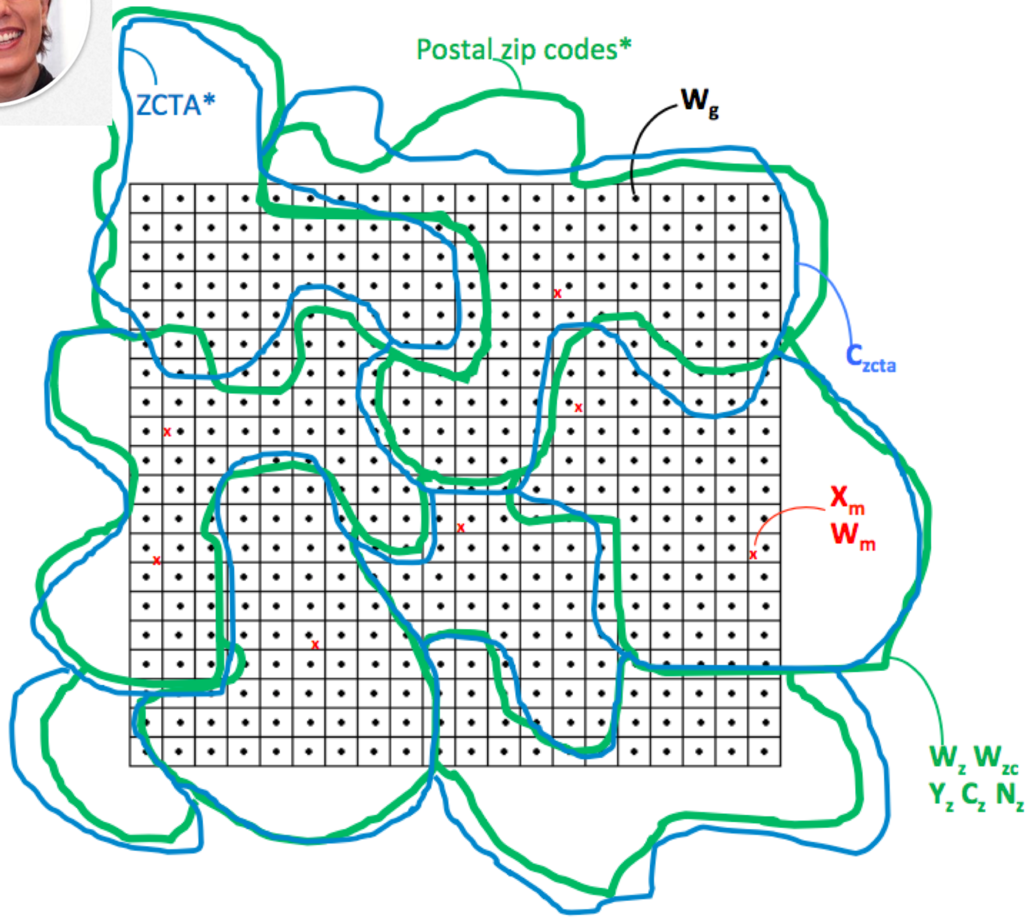
Hazard ratios and 95% confidence intervals are shown for an increase of 10 μg per cubic meter in $PM_{2.5}$ and an increase of 10 parts per billion (ppb) in ozone. Subgroup analyses were conducted by first restricting the population (e.g., considering only male enrollees). The same two-pollutant analysis (the main analysis) was then applied to each subgroup. Numeric results are presented in Tables S3 and S4 in the Supplementary Appendix. Dashed lines indicate the estimated hazard ratio for the overall population.

Methodological issues

- Evidence of causality
- Exposure measurement error
- Unmeasured confounding bias
- Discovery of heterogeneous subgroups
- Reproducibility


CONTROL GROUP
OUT OF CONTROL GROUP.

Exposure Error and Spatial Misalignment



Notation

m:	Monitor locations
X_m:	"true" measured $PM_{2.5}$ concentrations @ m
W_m:	Predicted $PM_{2.5}$ @ m
g:	Grid cells
W_g:	Predicted $PM_{2.5}$ @ g
z:	Zip codes
W_z:	Aggregated $PM_{2.5}$ @ z
W_{zc}:	Ordinal $PM_{2.5}$ @ z
Y_z:	Health outcomes @ z
N_z:	Medicare enrollees @ z
C_z:	Confounders @ z
C_{zcta}:	Confounders @ zcta
**	Exaggerated non-overlap

RANDOMIZATION INFERENCE FOR DISCOVERING EFFECT MODIFICATION IN AIR POLLUTION STUDIES

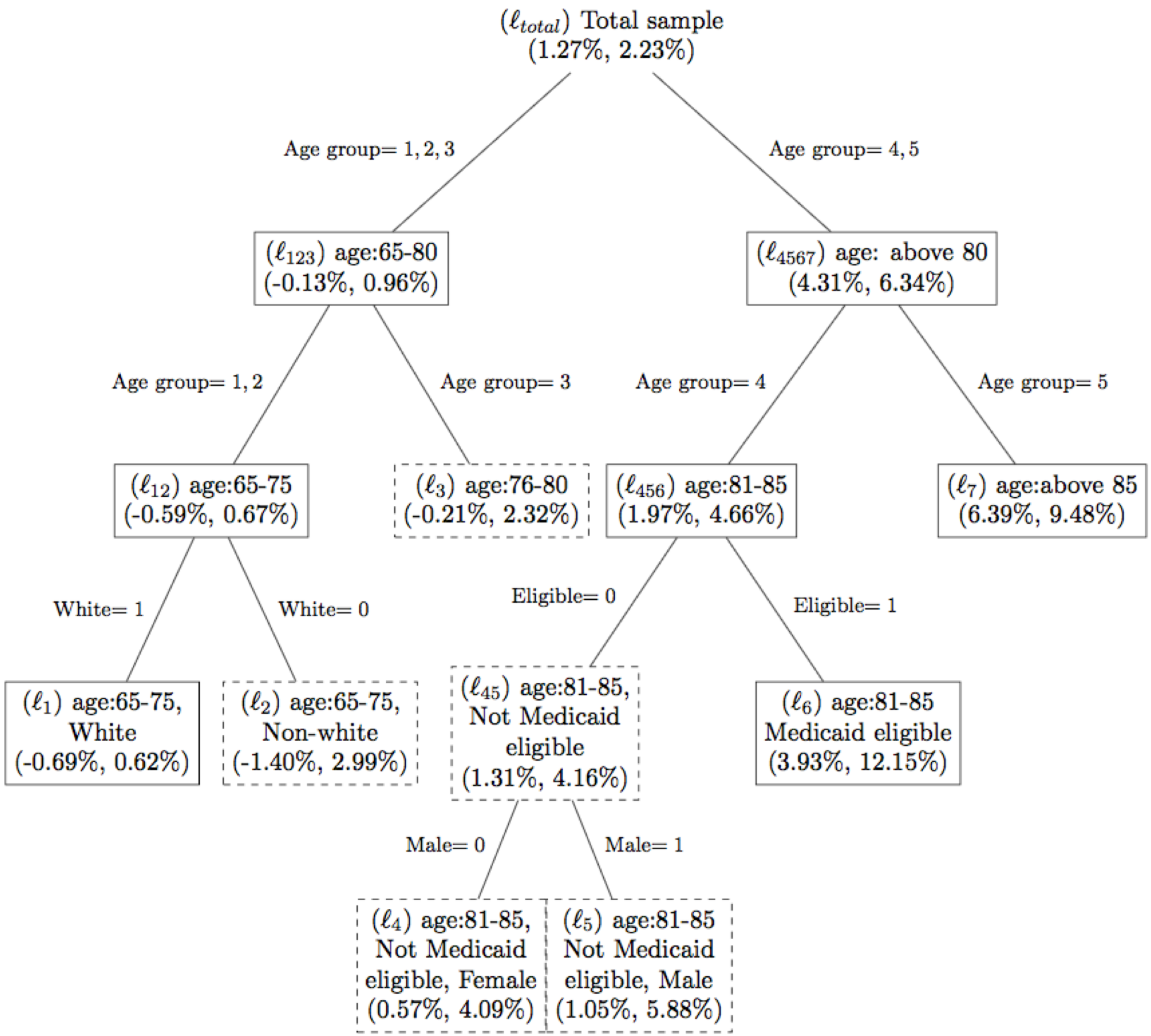
- We split the sample into two parts
- In the discovery step, our method considers machine learning techniques, especially tree algorithms (CART and CT), to uncover heterogeneous structures of treatment effects
- In the confirmation step, our method incorporates the discovered tree structures into a testing framework, and conducts hypothesis tests to confirm effect modification by combining with the CI method.

- Split + match + discover + test

Preliminary results: Medicare beneficiaries (N=1,612,414 individuals) who reside in the New England region from 2000 to 2006. Two year averages of fine particulate matter (PM2.5) from January 1, 2000 to December 31, 2001 is considered as exposures, and all-cause mortality is the outcome.

TABLE 2. Summary statistics and covariate balance before and after matching.

	Summary Statistics			Standardized Differences	
Covariates	Treated	Control (Before)	Control (After)	Before	After
Individual-level					
Male (%)	38.5	39.9	38.5	-0.02	0.00
White (%)	92.8	96.9	92.8	-0.19	0.00
Medicaid Eligible (%)	10.8	9.1	10.8	0.05	0.00
Age (Group, 1-5)	2.6	2.6	2.6	0.02	0.00
Age (65-107)	76.3	76.1	76.3	0.02	0.00
ZIP code-level					
Temperature	283.5	282.9	283.4	0.55	0.06
Humidity	76.1	76.9	76.1	-0.44	0.01
BMI (%)	26.1	26.3	26.1	-0.44	-0.06
Smoker Rate (%)	49.9	52.6	49.7	-0.72	0.07
Black Population (%)	6.2	3.2	6.0	0.33	0.03
Median Household Income	56.1	53.8	56.7	0.10	-0.03
Median Value of Housing	207.5	184.8	205.9	0.20	0.01
% Below Poverty Level	8.3	9.1	8.3	-0.09	0.01
% Below High School Education	30.6	30.1	30.2	0.03	0.03
% of Owner Occupied Housing	62.9	68.9	62.7	-0.33	0.01
Population Density (log-scale)	-6.9	-8.1	-7.0	0.89	0.06


(ℓ_{total}) Total sample
(1.27%, 2.23%)
Age group= 1, 2, 3
Age group= 4, 5
(ℓ_{123}) age:65-80
(-0.13%, 0.96%)
(ℓ_{4567}) age: above 80
(4.31%, 6.34%)
Age group= 1, 2
Age group= 3
Age group= 4
Age group= 5
(ℓ_{12}) age:65-75
(-0.59%, 0.67%)
(ℓ_3) age:76-80
(-0.21%, 2.32%)
(ℓ_{456}) age:81-85
(1.97%, 4.66%)
(ℓ_7) age:above 85
(6.39%, 9.48%)
White= 1
White= 0
Eligible= 0
Eligible= 1
(ℓ_1) age:65-75,
White
(-0.69%, 0.62%)
(ℓ_2) age:65-75,
Non-white
(-1.40%, 2.99%)
(ℓ_{45}) age:81-85,
Not Medicaid
eligible
(1.31%, 4.16%)
(ℓ_6) age:81-85
Medicaid eligible
(3.93%, 12.15%)
Male= 0
Male= 1
(ℓ_4) age:81-85,
Not Medicaid
eligible, Female
(0.57%, 4.09%)
(ℓ_5) age:81-85
Not Medicaid
eligible, Male
(1.05%, 5.88%)

The importance of open science and reproducible research

Replication vs reproducibility

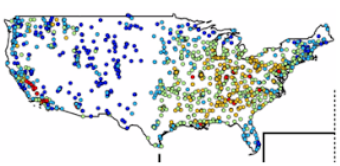
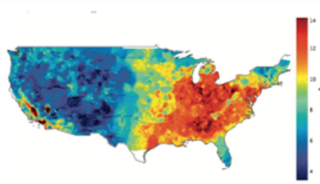
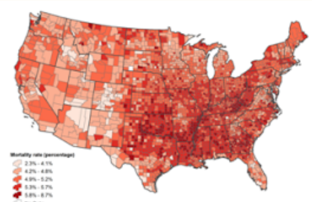
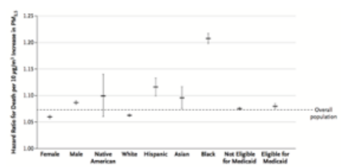
- A study is ***replicated*** when new data is collected and analyzed, independently, by a new set of investigators
- A study is ***reproduced*** when the same data is re-analyzed, independently by a new set of investigators

Challenges

- ***Scalability*** of computing and storage solutions. TB of data with private health information
- ***Privacy***. Secret Science Bill, HONEST Act

Open science and reproducible research


STEP 1	STEP 2	STEP 3	STEP 4	STEP 5
Data collection	Prediction model	Medicare claims	Statistical model	Reproducibility


30 TB of raw data from public sources processed to get 15TB of input data for the prediction model.	100 CPUs to calculate daily predictions 2000-2012 for $PM_{2.5}$ and ozone on a 1km x 1km grid.	Hundreds of GB of Medicare health outcomes processed to calculate ZIP-code level mortality.	632 jobs processed on 24TB of memory, 659 CPUs, 1.3 year of runtime.	Software codes and analytic datasets are hosted in digital data repositories.

Open science framework

- Steps 1 & 2: We rely on publicly-available data to generate TB of high-spatial resolution exposure predictions
- Step 3: By linking hundreds of GB of Medicare data to exposure predictions, we get TB of data that now contains protected health information
- Step 4: To perform estimation, we used 24TB of memory and 1.3 years of runtime on a secure cluster
- Step 5: Documented software codes are hosted in an open science digital platform

OSFHOME

My Projects Search Support Donate Francesca Dominici

Data Files Wiki Analytics Registrations Forks Contributors Settings

Private 0

NSAPH /

Data

Contributors: Christine Choirat, Qian Di, Francesca Dominici, Kirsten O'Brien, Chanmin Kim, marianthi-anna kioumourtzoglou, Joseph Antonelli, Petros Koutrakis, Brent Coull, Yan Wang, Georgia Papadogeorgou, Yun Wang, Kelvin Chi Cheung Fong, Liuhua Shi, Yara Abu Awad, Lingzhen Dai, Kevin Cummiskey, Antonella Zanobetti, Cory Zigler, Joel Schwartz, Yaguang Wei, Danielle Braun, Maayan Yitshak-Sade, Chen Chen, Ander Wilson, Xiao Wu, Rachel Nethery

Date created: 2016-10-20 10:41 PM | Last Updated: 2017-07-07 06:05 PM

Category: Data

Description: Add a brief description to your component

Wiki

Add important information, links, or images here to describe your project.

Files

Click on a storage provider or drag and drop to upload

Filter

Name	Modified
Data	
– OSF Storage	
– Datasets from Qian Di	
OSF Storage	

Citation

osf.io/j9wgq

Components

Add Component Link Projects

Datasets from Qian Di
Choirat, Di, Dominici & 4 more
11 contributions

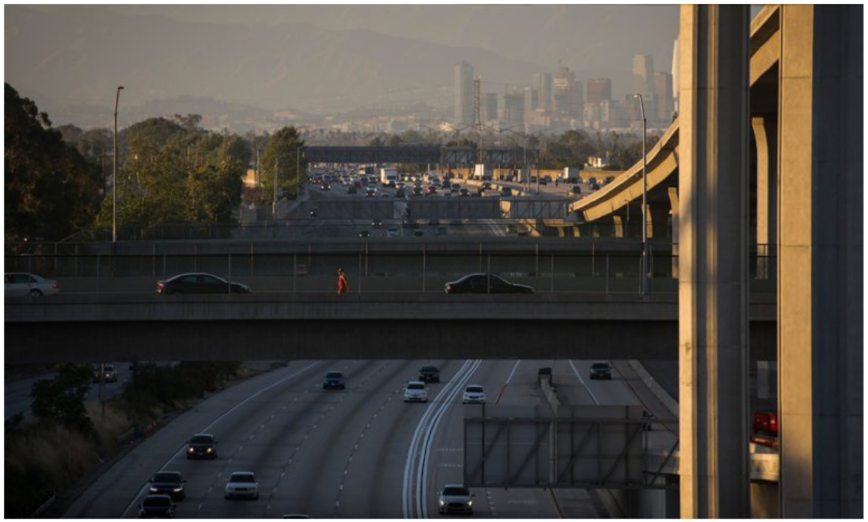
US postal zip codes
Choirat, Di, Dominici & 4 more
9 contributions

Datasets from Randall Martin's Group
Choirat, Kim, Cummiskey & 7 more

Even ‘Safe’ Pollution Levels Can Be Deadly

Leer en español

By NICHOLAS BAKALAR JUNE 28, 2017



PUBLIC HEALTH

U.S. Air Pollution Still Kills Thousands Every Year, Study Concludes

4:07

+ Queue

Download

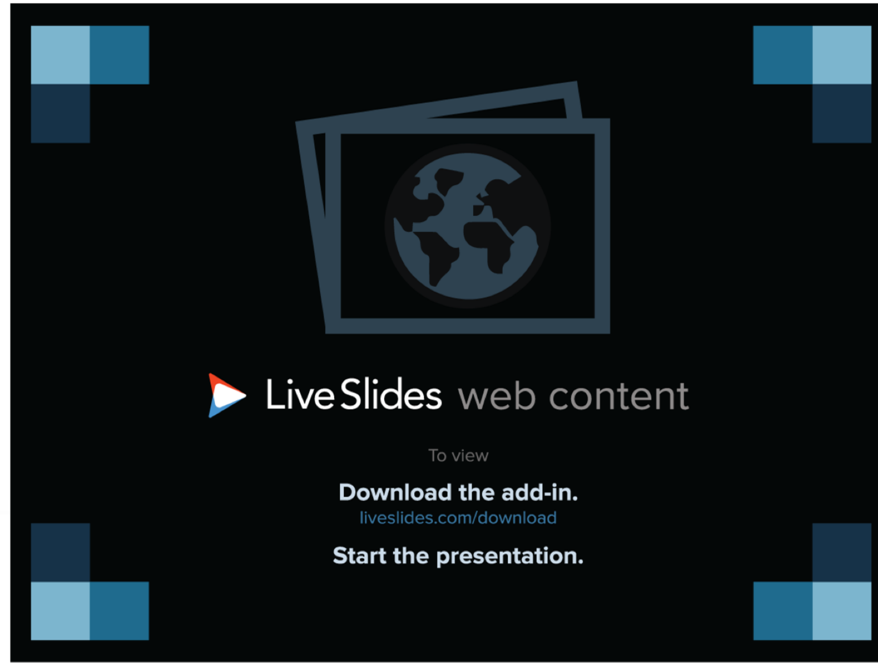
Embed

Transcript

June 28, 2017 · 5:01 PM ET
Heard on All Things Considered

ROB STEIN

A comprehensive study of air pollution in the U.S. finds it still kills thousands a year, and disproportionately affects poor people



Senator Cory Booker talking about the NEJM study at Hearing on the Nominations of Kathleen Hartnett White to be a Member of the Council on Environmental Quality. [https://player.vimeo.com/video/261362493]

LiveSlides web content

To view

Download the add-in.

liveslides.com/download

Start the presentation.

Energy and Environment

White House withdraws controversial nominee to head Council on Environmental Quality

By Brady Dennis and Juliet Eilperin February 4 Email the author

Kathleen Hartnett White, of the Texas Public Policy Foundation, arrives at Trump Tower on Nov. 28, 2016, in New York. (Drew Angerer/Getty Images)

Questions?

References

- *Makar M* et al (2017) Estimating the Causal Effect of Lowering Particulate Matter Levels below the United States Standards on Hospitalization and Death. *Epidemiology*
- *Di Q* et al (2017) Air Pollution and Mortality in the Medicare Population. *New England Journal of Medicine*
- *Dominici F, Zigler CM*. (2017). Best practices for gauging evidence of causality in air pollution epidemiology. *American Journal of Epidemiology*
- *Di Q* et al (2017) A Nationwide Case-crossover Study on Air Pollution and Mortality in the United States, 2000-2012, *Journal of American Medical Association,*
- *Wu et al* Causal inference in the context of error prone exposure: Air pollution and Mortality (submitted)
- *Lee et* al Randomization Inference for Discovery Effect Modification in Air Pollution studies (submitted)