HDSI | Harvard Data Science Initiative

HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH
Powerful ideas for a healthier world

How much evidence do you need? Data Science to Inform Environmental Policy During the COVID-19 Pandemic

Francesca Dominici, PhD
Professor of Biostatistics, Population Health and Data Science
Harvard T.H. Chan School of Public Health
Co-Director of the Harvard Data Science Initiative

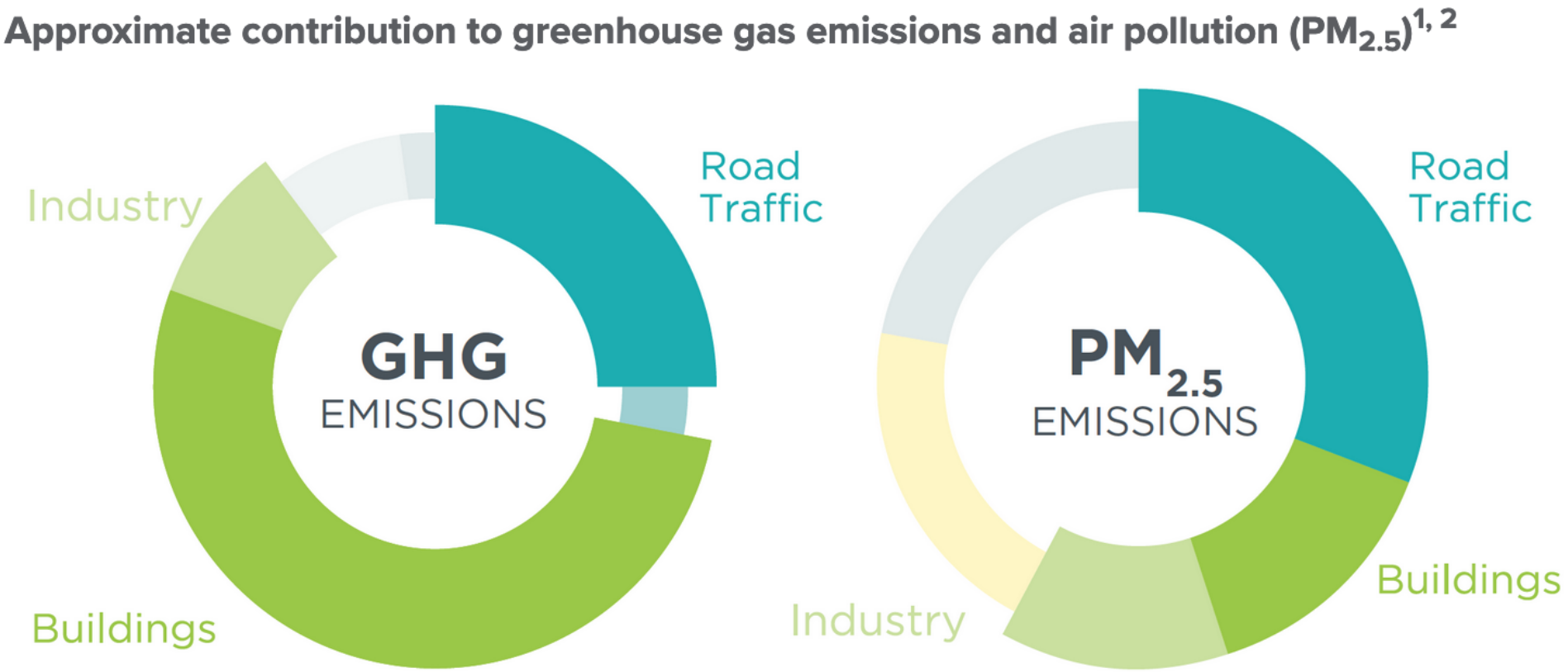
Key points

Satellite data, atmospheric chemistry models, and machine learning allow us to measure climate change related exposures (e.g wildfires) with unprecedented precision

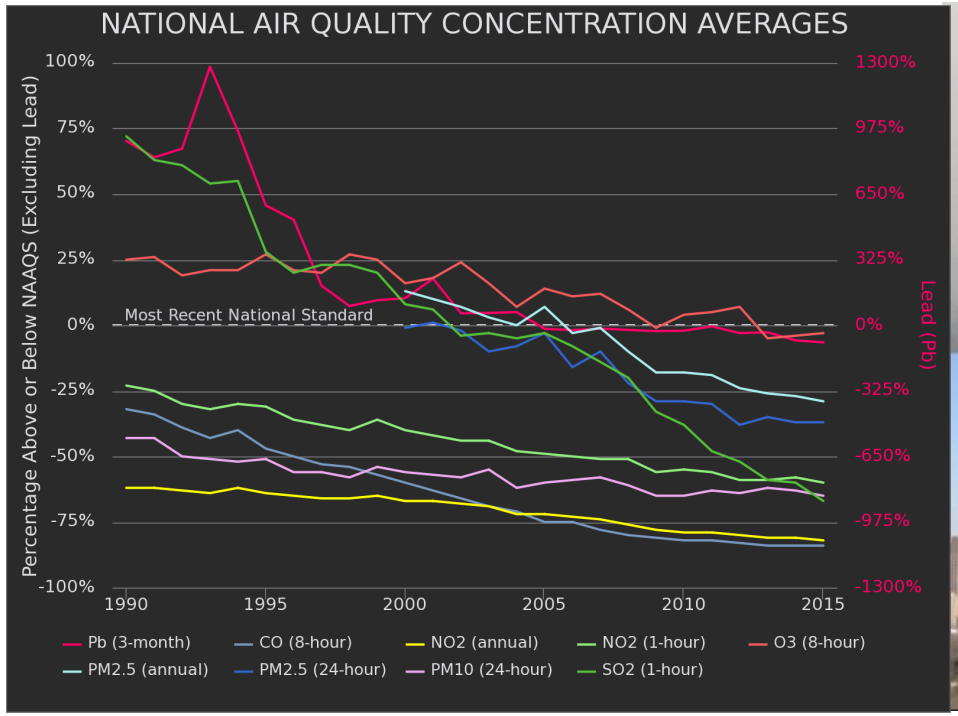
Methods for causal inference allows us to estimate health impacts from climate change related exposures and identify the most vulnerable

Emerging evidence of a link between air pollution, wildfires and COVID19 provide an additional sense of urgency

Climate change and air pollution share sources



DATA → DATA SCIENCE → EVIDENCE → POLICY CHANGE → CLEANER AIR



Scientific Questions

1. Is exposure to $PM_{2.5}$ below the NAAQS (12 µg/m^3) associated with an increase mortality risks?

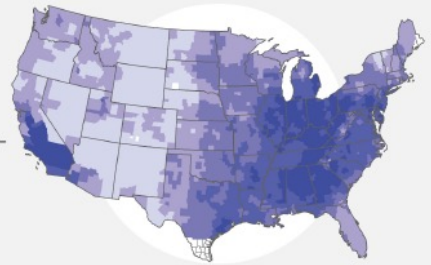
2. Are some populations at higher risk than others?

DATA

- All Medicare participants (n=67,682,479) in the continental United States from 2000 to 2016
- 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

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

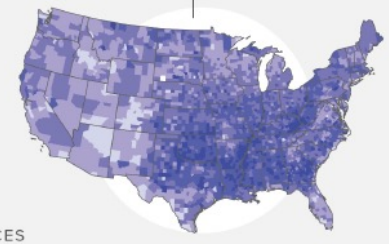
Traffic
Annual traffic counts and density from the Department of Transportation

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)



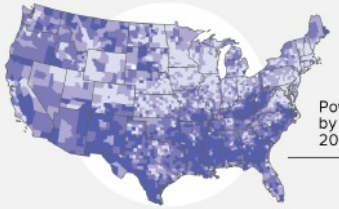
DATA SOURCES

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

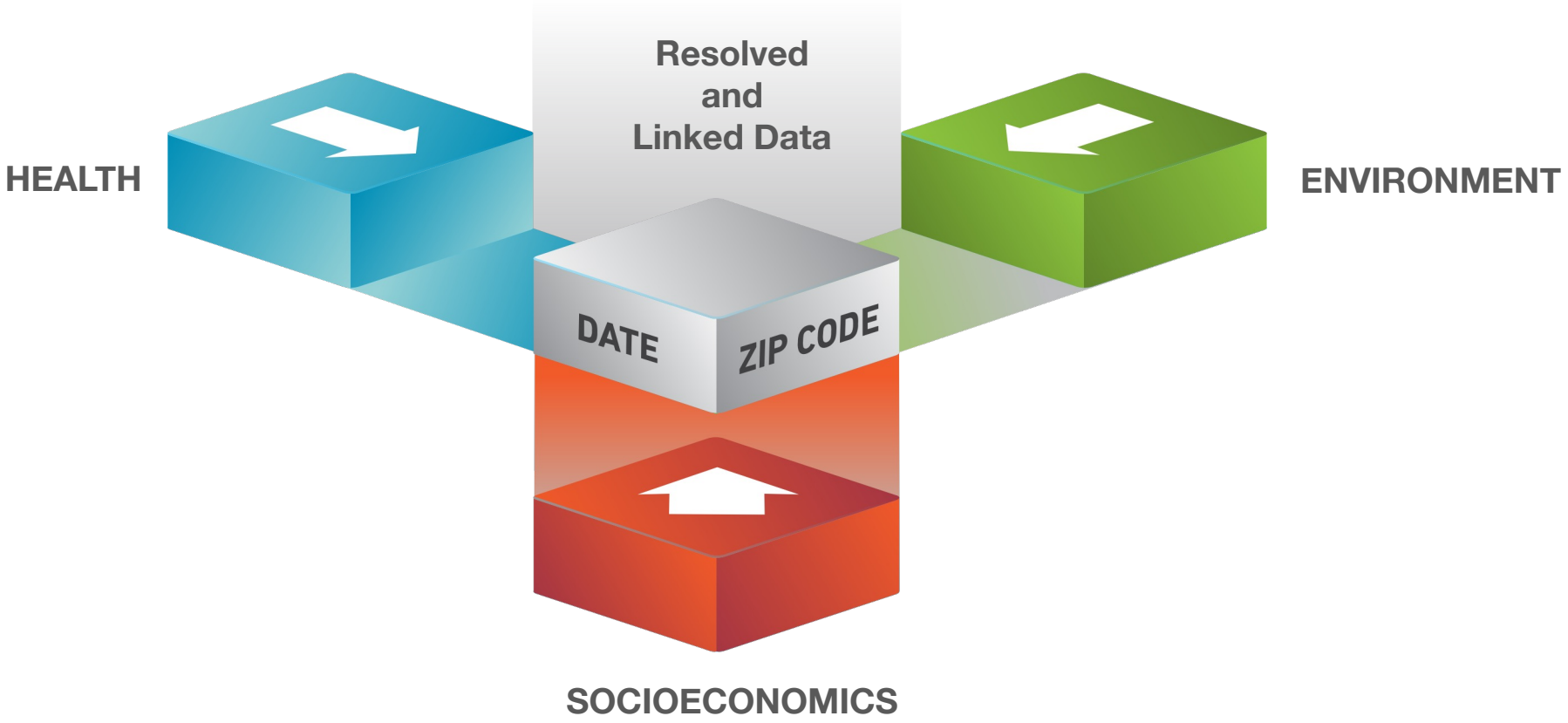
Individual demographics
Age, sex, race, ZIP code of residence

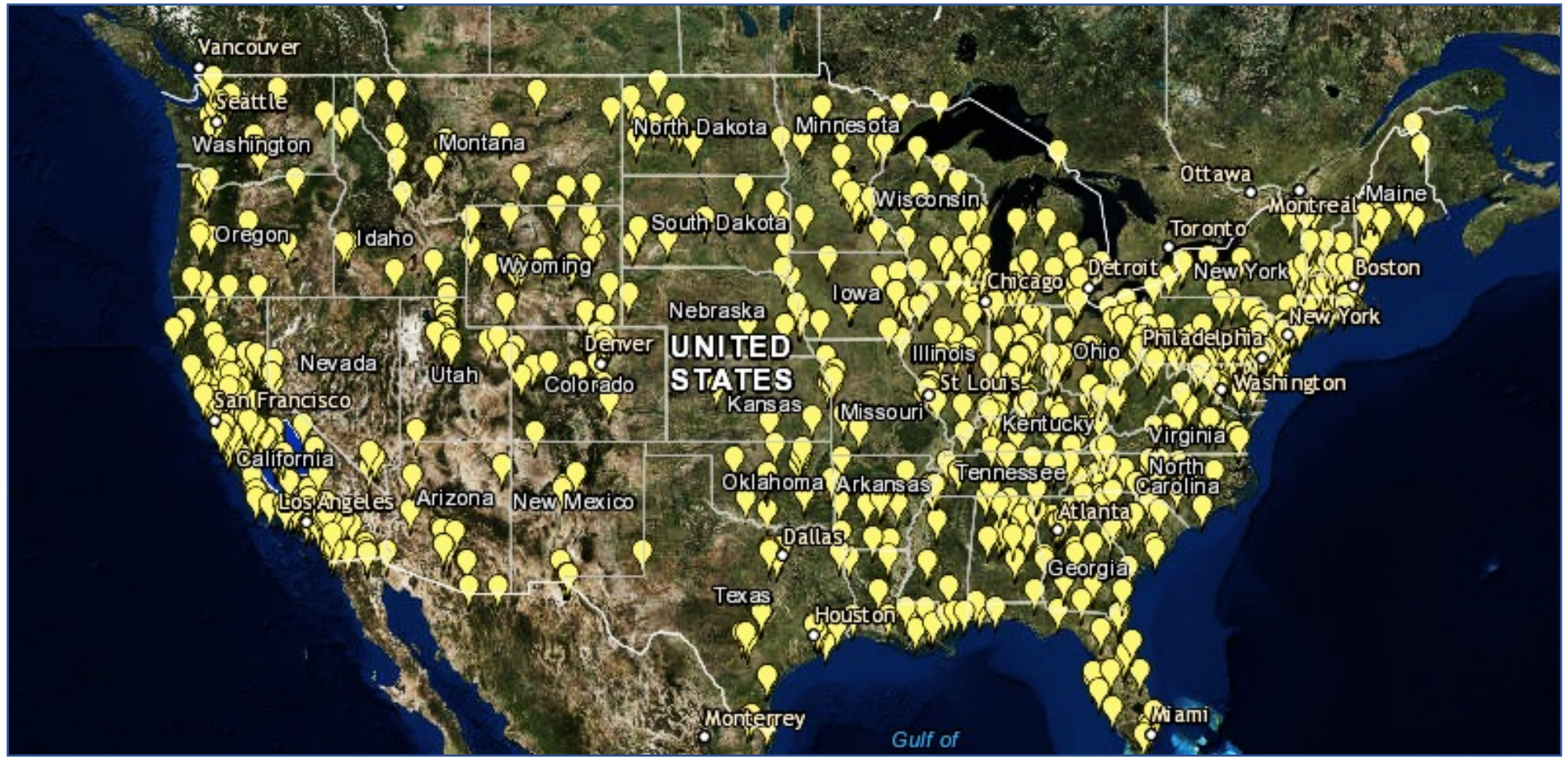
Individual medical history
Previous diagnoses, medications prescribed

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

County-level variables
Crime, smoking, BMI

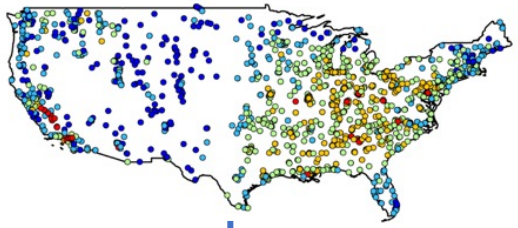
HARMONIZING AND INTEGRATING HETEROGENEOUS SOURCES OF DATA




Vancouver
Seattle
Washington
Montana
North Dakota
Minnesota
Ottawa
Montreal
Maine
Wisconsin
Oregon
Idaho
South Dakota
Toronto
Wyoming
Detroit
New York
Boston
Chicago
Iowa
Nebraska
New York
Denver
UNITED
STATES
Philadelphia
Nevada
Utah
Colorado
Illinois
Ohio
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

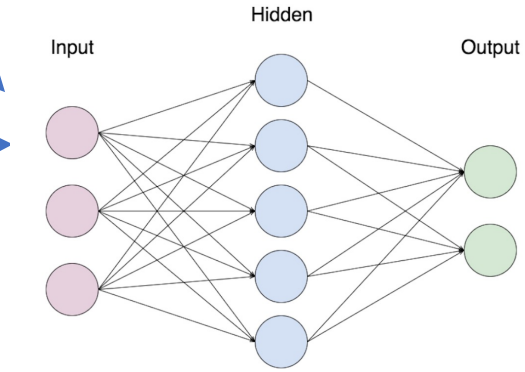
Satellite Imaging

$PM_{2.5}$ Monitor Data

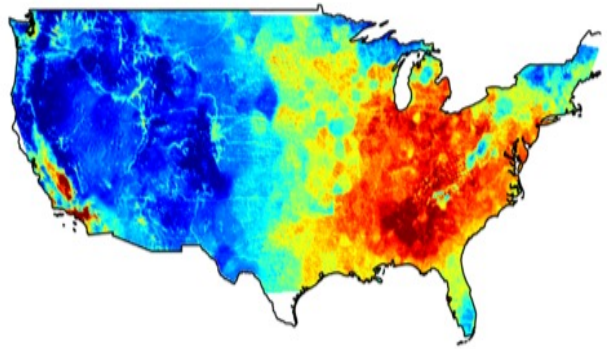


Joel Schwartz

Weather data



Daily 1km x 1km Estimates

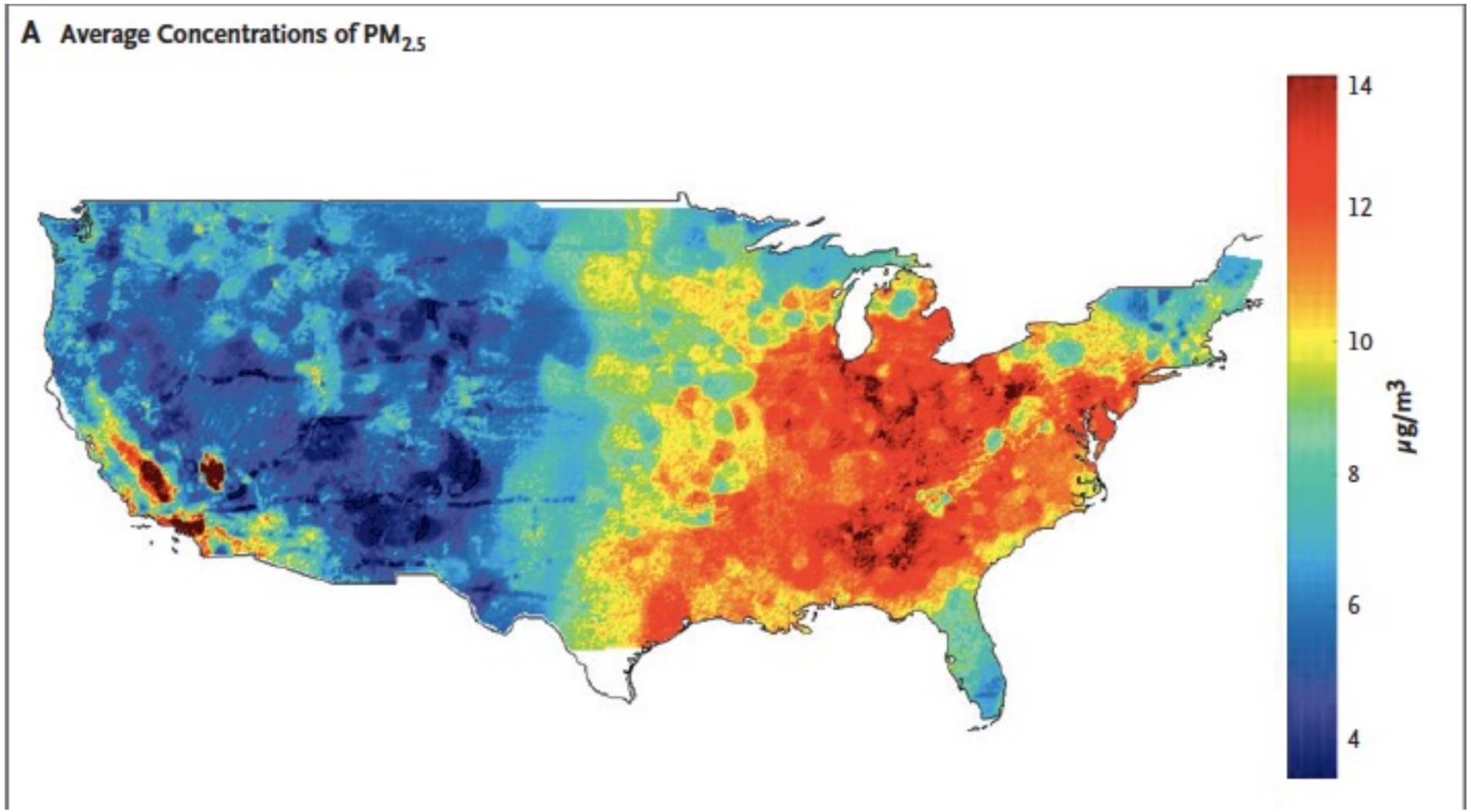


Land use data

Di Q et al. 2019. An ensemble-based model of PM2.5 concentration across the contiguous United States with high spatiotemporal resolution. Environ Int 130:104909, 10.1016/j.envint.2019.104909

Model	Pollutant	Space	Time	Period	Domain
AM3	O_3	0.5° × 0.5°	<daily[b]	2014-2016	global
AM3	$PM_{2.5}$	0.5° × 0.5°	daily	2014-2016	global
AM3	O_3	2° × 2°	<daily[b]	1980-2007	global
AM3	$PM_{2.5}$[a], NO_2	2° × 2°	daily	1980-2007	global
AM3	$PM_{2.5}$[a],O_3	2° × 2°	<daily[b]	2010-2014	global
AM3	NO_2	2° × 2°	monthly	2014-2016	global
GEOS-Chem	$PM_{2.5}$[a], O_3, NO_2	0.5° × 0.67°	<daily[b]	2004-2012	North America
GEOS-Chem	$PM_{2.5}$[a], O_3, NO_2	0.5° × 0.625°	<daily[b]	2013-2015	North America
GEOS-Chem	$PM_{2.5}$[a], O_3, NO_2	0.25° × 0.3125°	<daily[b]	select months /years	US
CMAQ	$PM_{2.5}$, O_3, NO_2	36 × 36 km		1990-2010	US
CMAQ	$PM_{2.5}$, O_3, NO_2	12 × 12 km		2007-2016	US
CMAQ	$PM_{2.5}$[a], O_3, NO_2	12 × 12 km	<daily[b]	2011	NE US
CMAQ (AMAD)	$PM_{2.5}$, O_3, NO_2	12 × 12 km	<daily[b]	2002-2012	NE US
HTAP models	$PM_{2.5}$, O_3, NO_2	36 × 36 km	<daily[b]	2010	US
CCMI models	$PM_{2.5}$, O_3, NO_2	1° × 1°	<daily[b]	1980-2010	global
CMAQ (AMAD) + AQS	O_3		MDA8[c]	2002-2012	NE[d] US
CMAQ (AMAD) + AQS	$PM_{2.5}$		daily	2002-2012	NE[d] US
MERRA-2 Aerosol reanalysis	$PM_{2.5}$[a]	0.5° × 0.625°	<daily[b]	1980-present	global
MACCRA	$PM_{2.5}$, O_3, NO_2	80 × 80 km	<daily[b]	2003-2012	global
CAMSiRA	$PM_{2.5}$[a], O_3, NO_2	110 × 110 km	<daily[b]	2003-2015	global
TCR-2	$PM_{2.5}$, O_3, NO_2	1.1° × 1.1°	<daily[b]	2005-2016	global
OMI + LUR	NO_2	100 × 100 m	annual	2011	global
OMI + model	NO_2	0.1° × 0.1°	monthly	2005 - 2016	global
OMI + LUR	NO_2	100 × 100 m	monthly	2000-2010	US
AOD + GEOS-Chem	$PM_{2.5}$[a]	0.1° × 0.1°	annual	1989-2013	global
AOD + LUR	$PM_{2.5}$	1 × 1 km	daily	2003-2011	SE[d] US
AOD + LUR + met + AQS	$PM_{2.5}$	1 × 1 km	daily	2000-2012	US
AOD + LUR + met + AQS	$PM_{2.5}$	10 × 10 km	daily	2000-2012	US
AOD + LUR + GEOS-Chem + met + AQS	$PM_{2.5}$	1 × 1 km	daily	2000-2014	US
OMI + LUR + GEOS-Chem + met + AQS	O_3	1 × 1 km	daily	2000-2014	US

Marianthi Kioumourtzoglou


A Average Concentrations of $PM_{2.5}$
14
12
10
8
6
4
μg/m³

ScienceAdvances

RESEARCH ARTICLES

Cite as: X. Wu *et al.*, *Sci. Adv*
10.1126/sciadv.aba5692 (2020).

Evaluating the impact of long-term exposure to fine particulate matter on mortality among the elderly

X. Wu,[1]† D. Braun,[1,2]† J. Schwartz,[3] M. A. Kioumourtzoglou,[4] F. Dominici[1]*

Study Population

- More than 68.5 million Medicare enrollees with $\geq$ 65 years old 2000-2016, who reside in 31,414 zip codes.
- Medicare claims data is an open cohort, including demographic information such as age, sex, race/ethnicity, date of death, and residential zip-code.
- A unique patient ID is assigned to each person to allow for tracking over time.

Table 1. Characteristics for the study cohorts

Variables	Entire Medicare Enrollees	Medicare Enrollees Exposed to $PM_{2.5}$ ≤12 $\mu g/m^3$
Number of individuals	68,503,979	38,366,800
Number of deaths	27,106,639	10,124,409
Total person-years	573,370,257	259,469,768
Median years of follow-up	8.0	8.0
Individual-level characteristics		

Data Sources

TABLE – Data Sources

	Source	Data
Exposure	Harvard University	1km $\times$ 1km $PM_{2.5}$ predictions
Meteorological	Google Earth Engine	4km $\times$ 4km temperature and relative humidity predictions
Confounders	Census	zip-code level socioeconomic status (SES) variables
	CDC	county-level behavioral risk factor variables
Health	CMS	mortality, individual-level characteristics

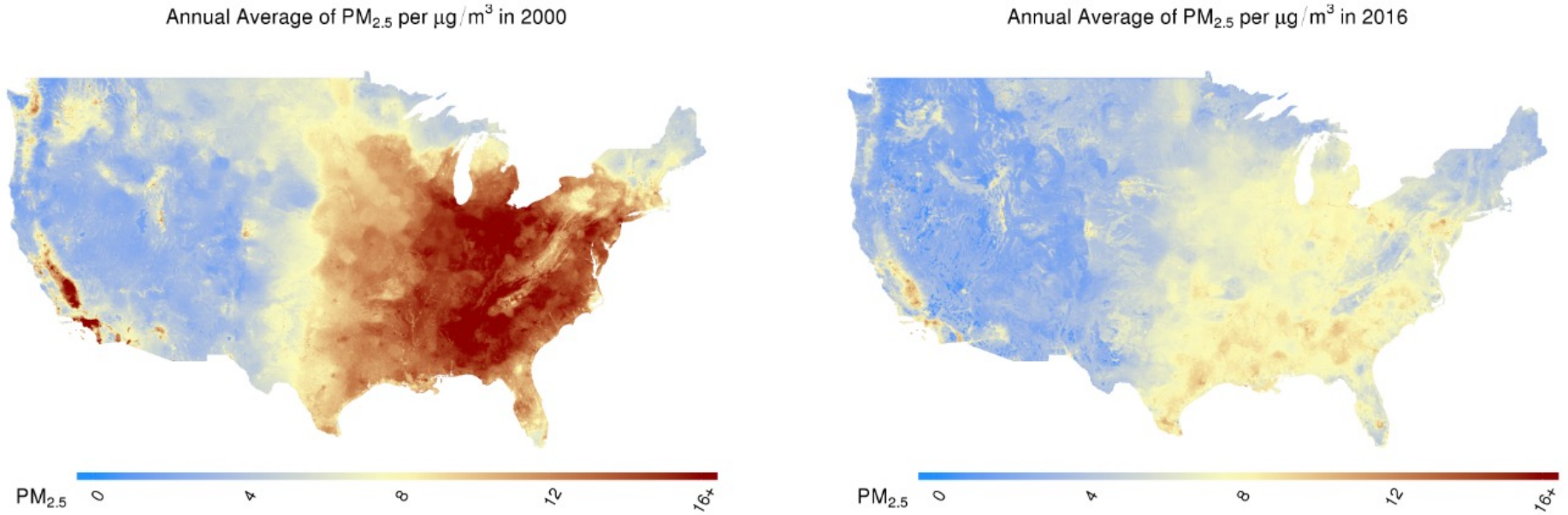


FIGURE – Annual average $PM_{2.5}$ concentrations in the continental United States for 2000 (top) and 2016 (bottom).

Potential Confounders

- Ten zip-code and county level confounders including zip-code level socioeconomic status (SES) from the 2000 and 2010 Census, county-level information from the Centers for Disease Control's Prevention's Behavioral Risk Factor Surveillance System.
- Four zip-code level meteorological variables : summer (June-September) and winter (December-February) average of 1) maximum daily temperatures and 2) relative humidity in each zip code obtained from Google Earth Engine.
- Two indicators for geographic region and calendar year.

Confounding is the largest threat to the validity of these studies

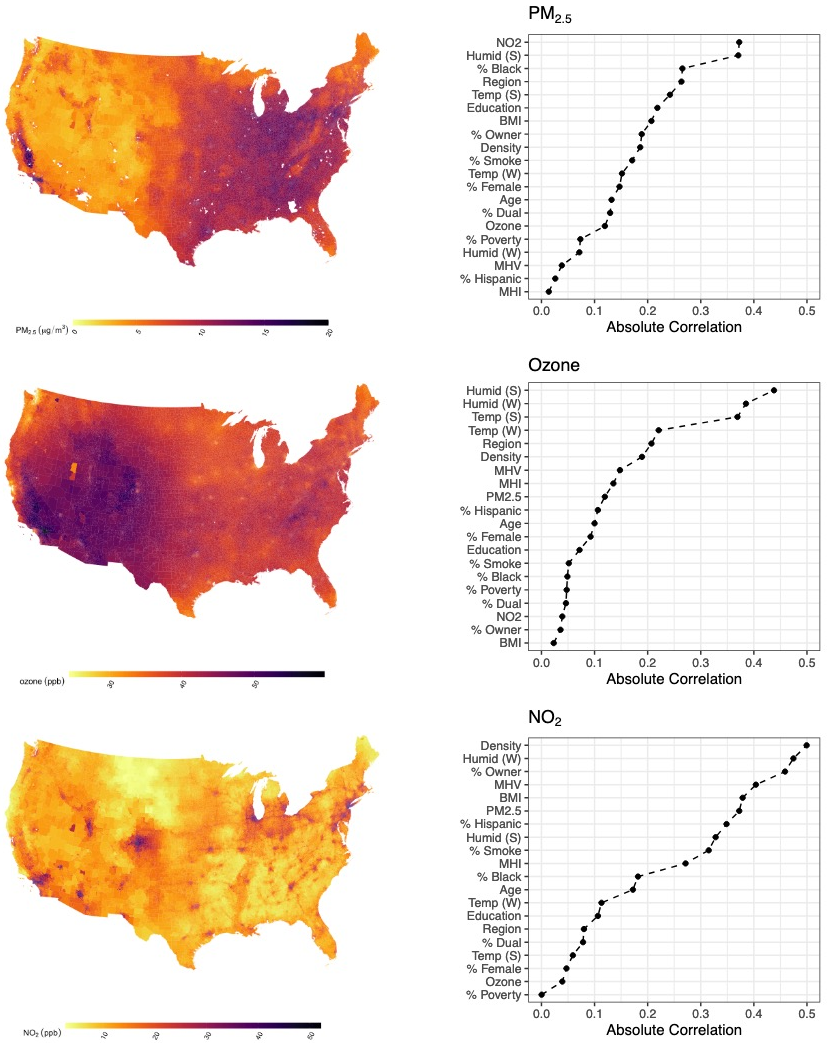


Figure 1: ***(Left)*** *ZIP-code-level average concentrations of $PM_{2.5}$, ozone and NO_2 across 2000 to 2016. Areas in white are those without $PM_{2.5}$ measurements.* ***(Right)*** *Correlations between each of the three pollutants and other variables. Age = average age at enrollment; Education = proportion of below high school education; Humid (S) and (W) = average summer and winter humidity; Region = census region; Temp (S) and (W) = average summer and winter temperature; Density = population density; % Owner = % owner-occupied housing; % Dual = proportion of dual eligibility of Medicare and Medicaid; MHV = median home value; MHI = median household income; BMI = average body mass index.*

Statistical Analysis

We fit five distinct statistical models to estimate the causal relationship between long-term $PM_{2.5}$ exposure and our outcome of interest, all-cause mortality among the elderly.

- Traditional approaches
  1. Cox Proportional Hazard Approach
  2. Poisson Regression Approach
- Causal Inference Approaches using Generalized Propensity Score (GPS)
  1. Matching Approach
  2. Weighting Approach
  3. Adjustment Approach

Fundamental idea of the GPS matching

- We want to construct matched datasets that approximate a randomized experiment as closely as possible by achieving good covariate balance.
- In the continuous exposure setting, the challenge is that it is unlikely that two units will have the exact same level of exposure
- Therefore, we proposed an approach that jointly matches on both the estimated GPS and exposure values.
- The closeness of exposure level guarantees that the matched unit is a valid representation of observations for a particular exposure level, whereas the closeness of GPS ensures that we are properly adjusting for confounding.

- Code available ! https://github.com/wxwx1993/GPSmatching
- Paper under review and available https://arxiv.org/abs/1812.06575

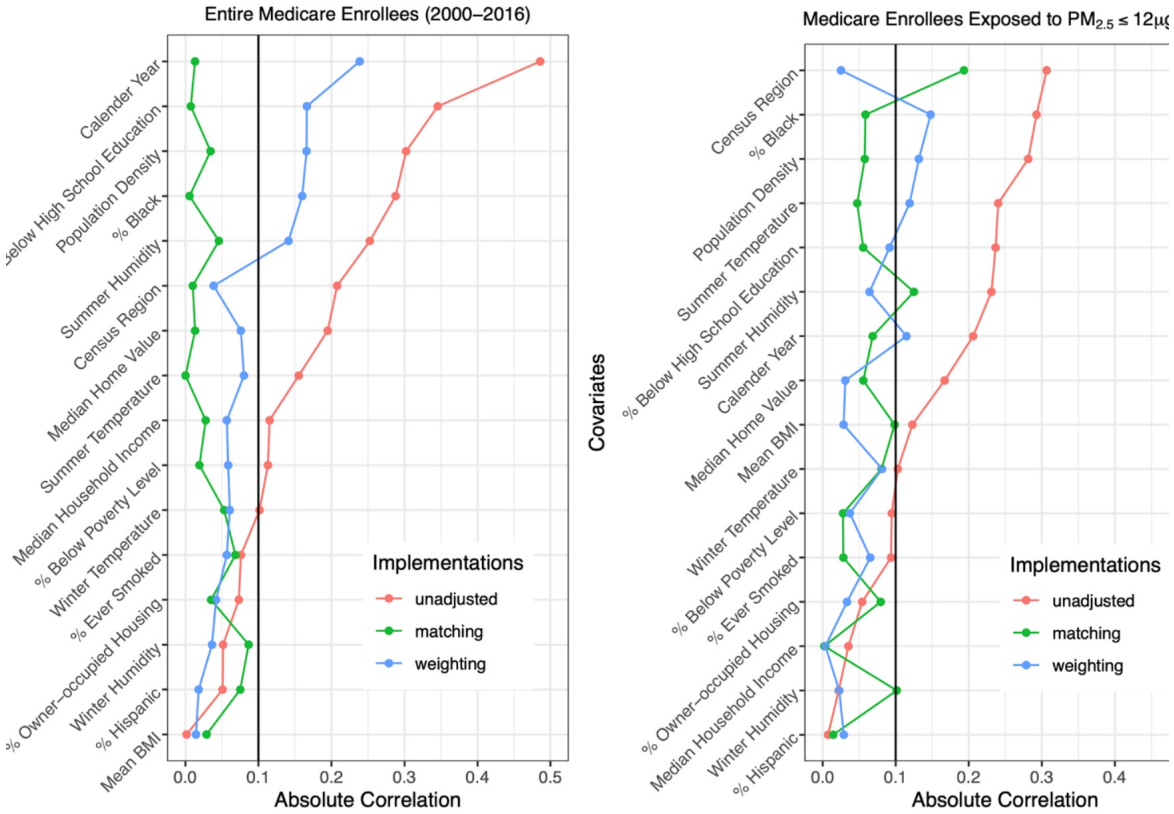


Fig. 2. Mean absolute correlation (AC) for Unadjusted, Weighted, and Matched Populations. Mean AC was smaller than 0.1 using causal inference GPS methods (matching and weighting). AC values <0.1 indicate good covariate balance, strengthening the interpretability and validity of our analyses as providing evidence of causality.

The causal inference framework lends itself to the evaluation of covariate balance for measured confounders. An absolute correlation (AC), with values <0.1 indicating high quality recovering randomized experiments.

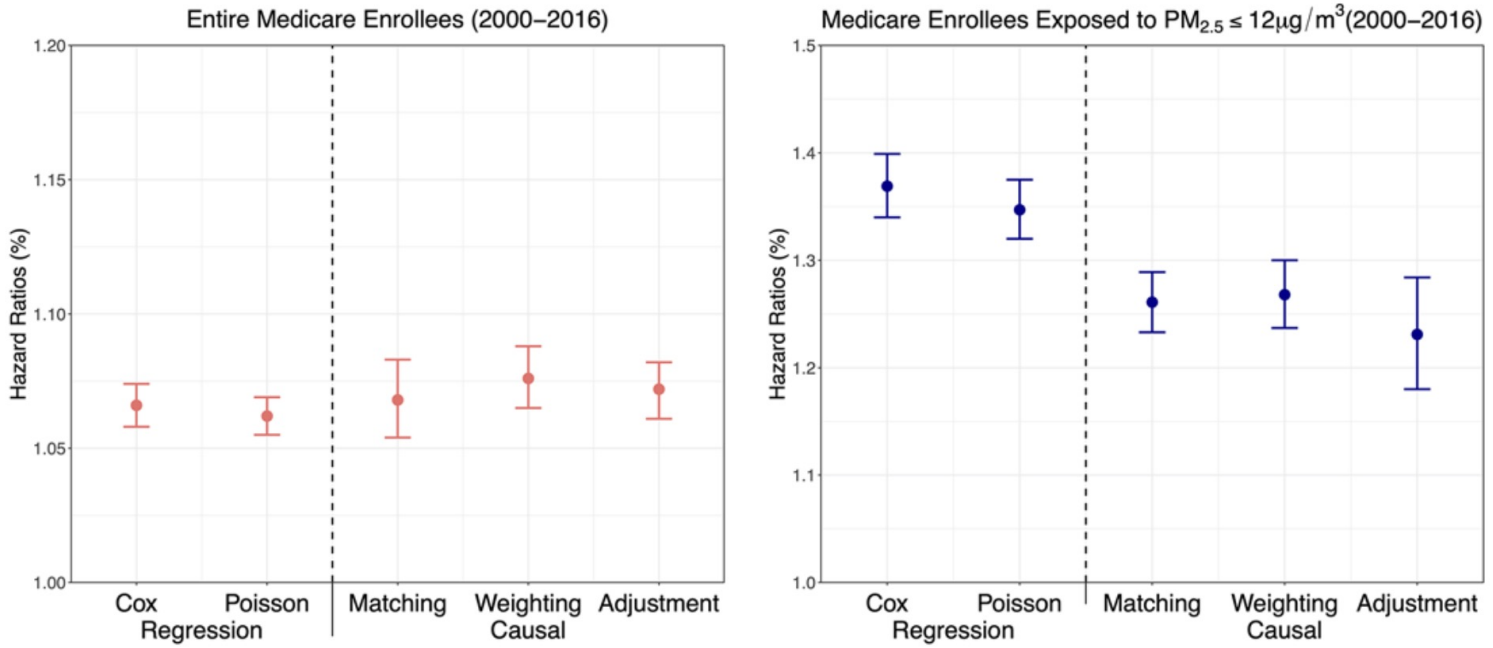


Fig. 3. Hazard Ratios (HR) and 95% Confidence Intervals (CIs). The estimated HRs were obtained under five different statistical approaches (two traditional approaches and three causal inference approaches). HRs were adjusted by 10 potential confounders, four meteorological variables, geographic region, and year.

Using five distinct statistical approaches, we found that a decrease of 10 μg/m3 PM2.5 leads to a statistically significant 6%–7% decrease in mortality risk.

Based on these models, lowering the air quality standard to 10 μg/m3 would save 143,257 lives (95% 30 confidence interval 115,581–170,645) in one decade

- Experts said that efforts to curb pollution by reducing fossil fuel use would provide a double benefit, in both improving public health conditions and bringing down climate-warming emissions.

WHO slashes guideline limits on air pollution from fossil fuels

Level for the most damaging tiny particles is halved, reflecting new evidence of deadly harm

▲ A coal-burning power plant. The WHO says at least 7 million people a year are killed by air pollution.

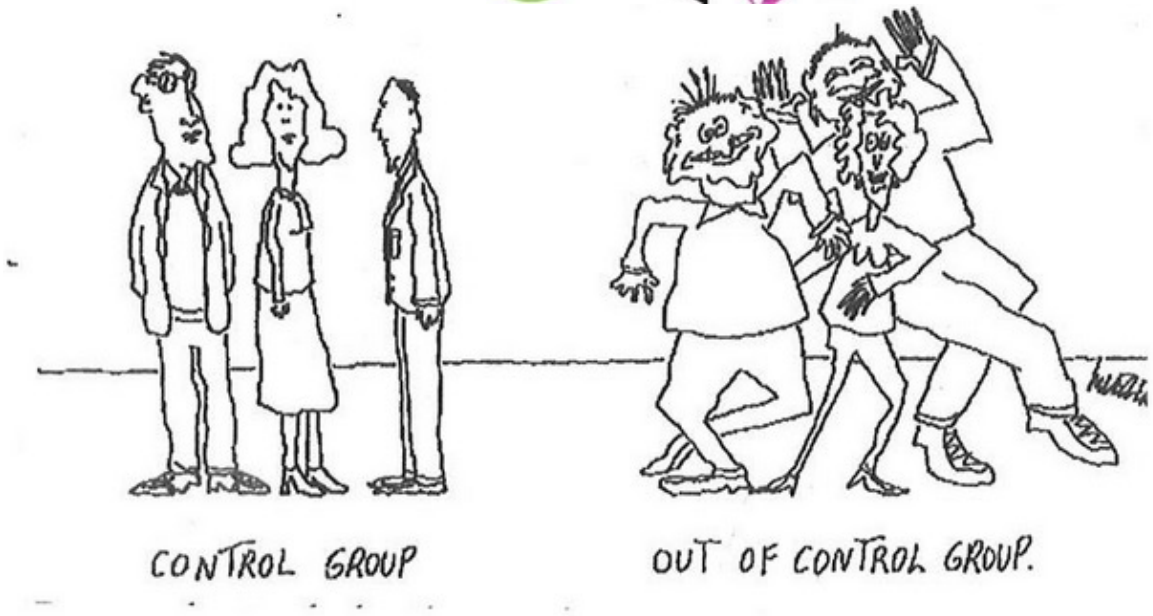
EPA to Reexamine Health Standards for Harmful Soot that Previous Administration Left Unchanged

06/10/2021

Contact Information:
EPA Press Office (press@epa.gov)

The strong body of scientific evidence shows that long- and short-term exposures to fine particles (PM2.5) can harm people's health, leading to heart attacks, asthma attacks, and premature death. Large segments of the U.S. population, including children, people with heart or lung conditions, and people of color, are at risk of health effects from PM2.5. In addition, a number of recent studies have examined relationships between COVID and air pollutants, including PM, and potential health implications. While some PM is emitted directly from sources such as construction sites, unpaved roads, fields, smokestacks or fires, most particles form in the atmosphere as a result of complex reactions of chemicals such as sulfur dioxide and nitrogen oxides, which are pollutants emitted from power plants, industrial facilities and vehicles.

From Hypothesis first to Data First
From Randomization to Observation



Unresolved Data Science Challenges

- Uncertainty quantification and propagation
- Causality
  - Unmeasured confounding
  - Estimation of exposure response curve
  - Heterogeneity
- Reproducibility

Applications and Case Studies

Evaluation of the Health Impacts of the 1990 Clean Air Act Amendments Using Causal Inference and Machine Learning

Rachel C. Nethery, Fabrizia Mealli, Jason D. Sacks & Francesca Dominici

Received 21 Aug 2019, Accepted 20 Jul 2020, Accepted author version posted online: 31 Jul 2020, Published online: 16 Sep 2020

Download citation https://doi.org/10.1080/01621459.2020.1803883 Check for updates

Applications and Case Studies

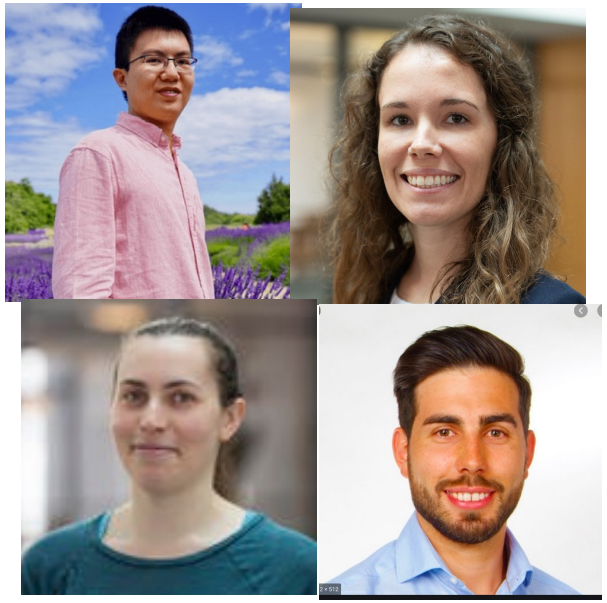
Estimating the Effects of Fine Particulate Matter on 432 Cardiovascular Diseases Using Multi-Outcome Regression With Tree-Structured Shrinkage

Emma G. Thomas, Lorenzo Trippa, Giovanni Parmigiani & Francesca Dominici

Received 17 Jan 2019, Accepted 22 Jan 2020, Accepted author version posted online: 24 Feb 2020, Published online: 26 Feb 2020

Download citation https://doi.org/10.1080/01621459.2020.1722134 Check for updates

Methodological contributions in Causal Inference and Machine Learning



Matching on Generalized Propensity Scores with Continuous Exposures

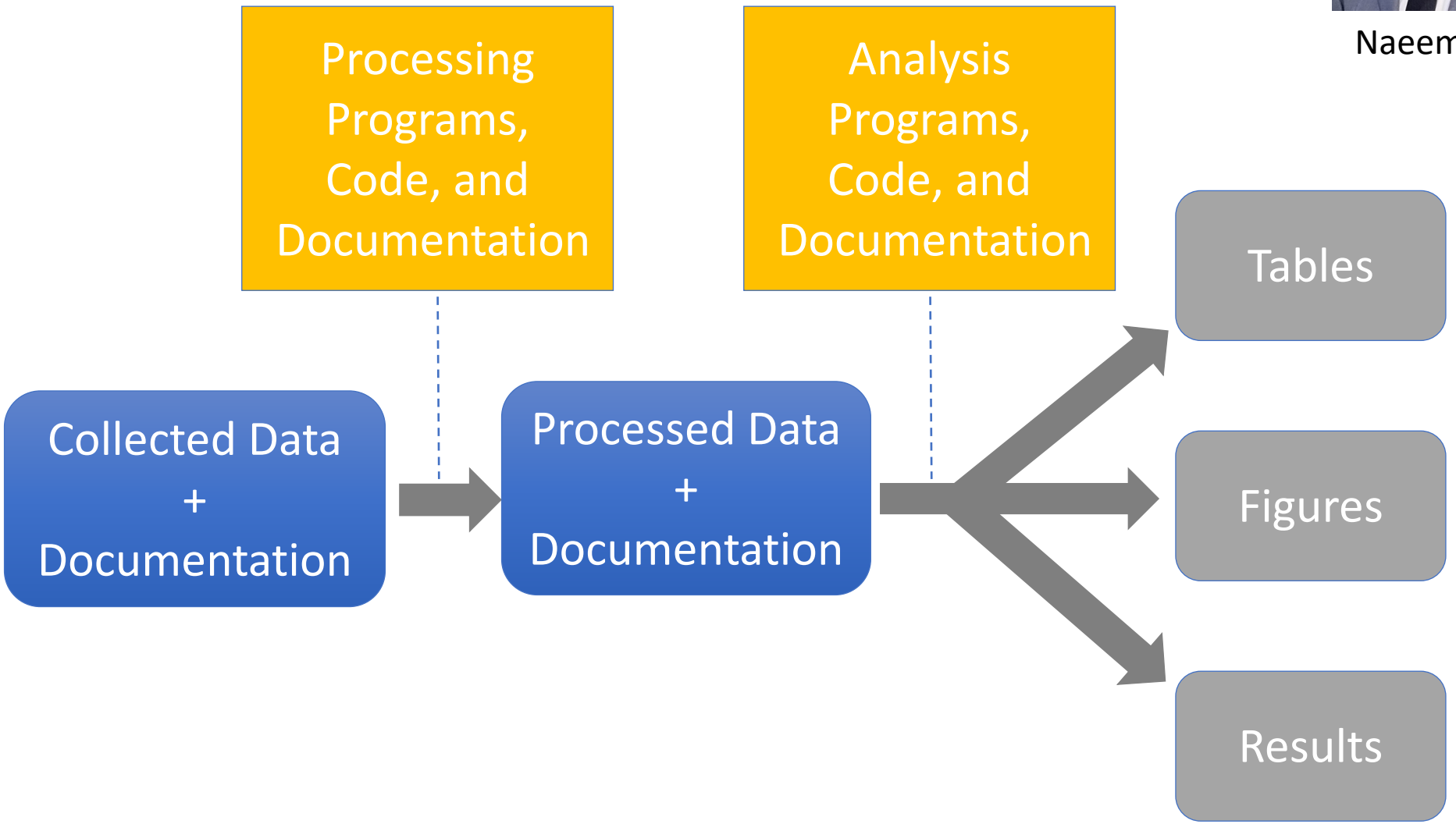
Xiao Wu, Fabrizia Mealli, Marianthi-Anna Kioumourtzoglou, Francesca Dominici, Danielle Braun

Causal Rule Ensemble: Interpretable Inference of Heterogeneous Treatment Effects

Kwonsang Lee, Falco J. Bargagli-Stoffi, Francesca Dominici

Reproducible workflows

Naeem Khoshnevis



National Studies on Air Pollution and Health "NSAPH"

- **8 PIs from other institutions:**
  - Columbia University
  - Yale University
  - Boston University
  - The University of Texas at Austin
  - University of Florida
  - University of British Columbia
  - University of Rochester

- **15 Post-Doctoral Fellows**
- **16 PhD students**
- **5 Master's students**
- **7 Undergraduate students**

Leila Kamareddine

NSAPH Team Members

Francesca Dominici

Michelle Bell

Joel Schwartz

Danielle Braun

Marianthi Kioumourtzoglou

Rachel Nethery

David Christiani

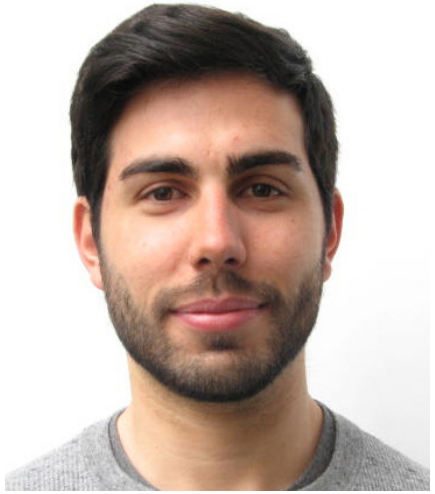
Gregory Wellenius

Antonella Zanobetti

Tanujit Dey

NSAPH Team Members ctd’

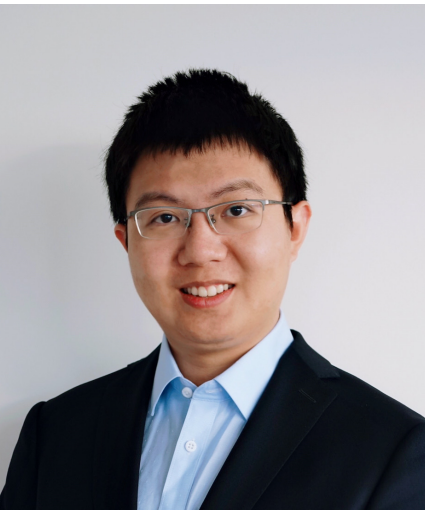
Kate Weinberger



Falco J. Bargagli Stoffi

Naeem Khoshnevis

Leila Kamareddine



Xiao Wu

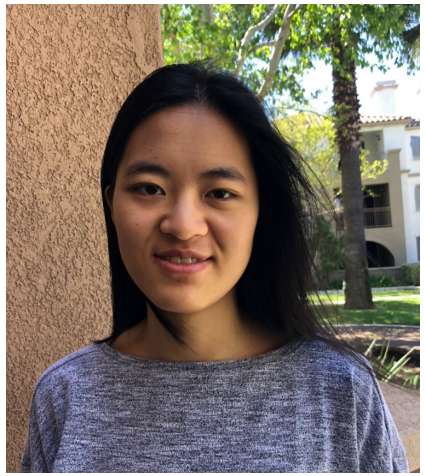
Kevin Josey

Robbie Parks

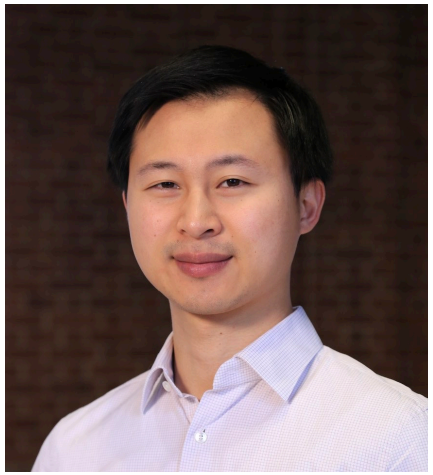
Mahdieh Danesh Yazdi

NSAPH Team Members ctd’ & many more!

Kate Burrows

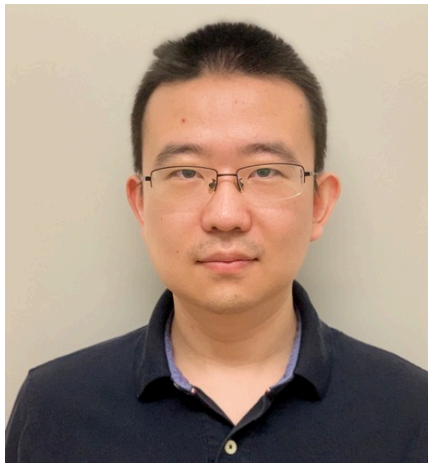


Xiaodan Zhou

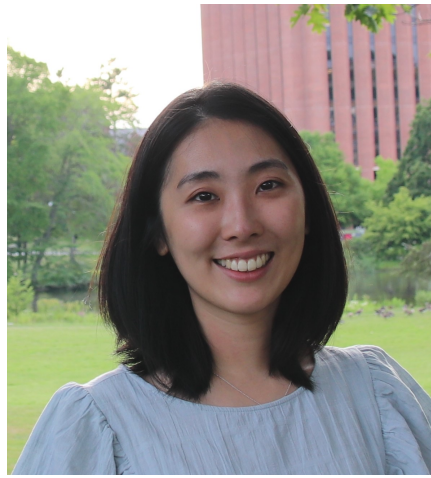


Kelvin Fong

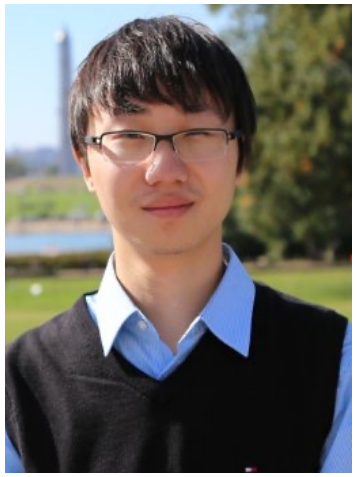
Kaela Nelson



Boyu Ren



Jenny Lee



Yaguang Wei

Ellen Considine

Yue Yang

Michael Cork

Abdulrahman Jbaily

Kevin Lee Chen

Jeremiah Jones

Conclusions

- The steps needed to mitigate climate change in the future are substantially the same as those needed to reduce the burden of death and disability due to air pollution in the present — **cut back on burning fossil fuels and biomass**
- In the meantime, machine learning and data science allow us to measure exposure and pinpoint susceptibility and vulnerability
- Methods for causal inference allow us to better disentangle causes from confounders, especially in the context of natural disasters (Nobel prize in economics)

Looking ahead: a data science perspective

- Develop new research data platforms
- Link spatial-temporal data on meteorology, climate, air pollution, land use, and satellite sensor readings to understand and quantify linkages to health indicators
- Understand the distribution of effects on disadvantaged portions of the population.