CAUSAL EFFECTS OF PM2.5 POLLUTION EXPOSURE ON PEDIATRIC HEALTH OUTCOMES

Trevor Johnson; Matt Lyons; Anand Patel, MS; Michelle Shen; Cornelia Ilin, PhD

Berkeley SCHOOL OF INFORMATION

INTRODUCTION

Air pollution is a pervasive, global issue worsening due to climate change, increasing population, deforestation, and other causes. Among different types of air pollution, PM2.5 pollution, which consists of fine particulate matter smaller than 2.5 microns, poses significant health effects, as it can enter the bloodstream.

Prior studies have found that proximity to sources of pollution is associated with increased risk of childhood cancer, cardiovascular and respiratory illnesses, end-stage renal disease, and diabetes (Brender, Maantay, and Chakraborty, 2011), but the breadth and extent of contemporaneous and long-term impacts of PM2.5 exposure on hematopoietic cancers, type 1 diabetes, and vasculitis are not well understood in children.

We develop a causal model for estimating the effects of pollution exposure on health using variation in local air pollution driven by wind speed and direction, as well as proximity to schools in the state of California over a 15-year period.

OBJECTIVES

We aim to examine the link between PM2.5 pollution exposure and adverse pediatric health outcomes throughout the state of California from 2000 to 2017. More specifically, our aims are as follows:

Aim 1: Issues of air quality disproportionately affect children who live in less wealthy areas. To understand this issue better, the first goal is to build unique high-quality data of PM2.5 pollution exposure at every school in CA.

Aim 2: Establish an empirical relationship between PM2.5 levels recorded at schools in CA and being downwind from key point sources of pollution.

Aim 3: Estimate the causal effect of a school being located downwind from key pollution sources on emergency room visits and hospitalization rates for hematopoietic cancers, type 1 diabetes, and pediatric vasculitis conditions among patients 0–19 years old.

Aim 4: Conduct a counterfactual study to compare and quantify the benefits of various air pollution mitigation strategies.

Aim 5: Develop a model to predict future emergency room visits and hospitalization rates for our conditions among patients 0–19 years old, using demographics, pollution properties, and time-series trends in addition to wind.

DATASETS

We assembled the following data elements from the years 2000–2017:

- 13,297 **school centroids** in 1,391 **California zip codes** (Cal. Dept. of Ed.)
- 56.2 million **hospital stays with diagnoses** (CDPH, OSHPD)
- **PM2.5 readings** at each zip code/month/year combination (WUSTL ACAG)
- **Wind speed and direction** for each zip code/month/year (NASA MERRA-2)
- Zip-code/year level **income and SES estimates** (CA Franchise Tax Board)
- 7,155 **PM2.5 pollution point sources** (EPA National Emissions Inventories)
- **Elevations** at all schools and point sources (USGS)
- **Population and demographics** by zip code/year (US Census Bureau)
- **Temperature** for each month/year at each location (NOAA via Meteostat)

METHODS

Data Sources: We obtain inpatient and emergency health outcomes data from the California Department of Public Health (CDPH) and the California Office of Statewide Health Planning and Development (OSHPD) covering nearly 7,000 California-licensed healthcare facilities. These data are at the individual level, and include residential zip code information along with up to 21 ICD-9/ICD-10 diagnosis codes for each visit.

We next link health outcomes to locations of schools, locations of EPA-recognized major air pollution sites, wind speed and direction, and PM2.5 readings for each month and year in our data collection period (Fig 1). The distribution of the amount of PM2.5 emitted by the pollution point sources is right-skewed with a median PM2.5 output of 4.1 tons per year and standard deviation of 54.9. Average wind speeds have a median value of 1.0 meters per second with a standard deviation of 0.8. The resulting PM2.5 readings in each zip code have a median of 8.9 tons per year with a standard deviation of 6.0. We also incorporate additional data elements from several sources (listed at bottom-left in Datasets).

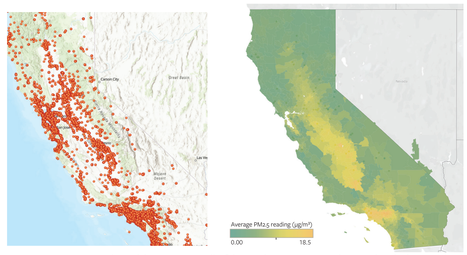


Fig 1. School locations (left) and PM2.5 readings (right)

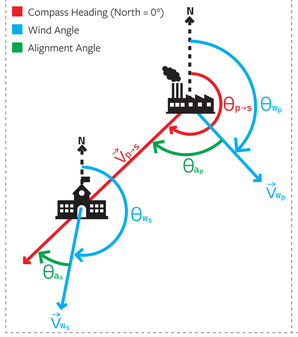
Data Cleaning: For numeric features with a lower sampling frequency than is required for our analysis (e.g. census populations), we use linear interpolation to fill in the gaps. For nearest pollution sources, we use the most recently-available data (e.g. 2004 point sources would be based on the 2002 National Emissions Inventory). We use GeoPy and the school latitude and longitude to retrieve any missing zip codes. We filter schools based on open/close dates, and we drop rows where crucial data elements are not available (e.g. PM2.5, wind, census population).

For causal analysis, we create measures of the distance between each school location and key point sources of pollution. We then combine our distance measures with wind direction data to determine if a school is up- or downwind from the point source. The rationale for using wind direction is to mimic an experimental study by using observational data. Wind direction acts as a natural experiment that divides schools into treatment and control groups.

Instrument Design: To transform the wind direction into a usable instrument, we express the heading from the nearest pollution source to its associated school as an angle from north (Fig 2, shown in red). Next, we express wind speed data in the same way (Fig 2, shown in blue). Finally, we take the average of the difference between these angles to account for different wind directions at the school and pollution site (Fig 2, shown in green). This alignment angle $\theta_{\text{downstream}}{}^{\ddagger}$ is 0° when the wind is blowing directly from the pollution site toward the school and 180° when it is blowing in the opposite direction.

Linking Datasets: The availability of major sources of pollution, school locations, PM2.5, and wind data varies across zip codes included in the analysis. We merge these data based on zip code, month, and year, to form a single longitudinal (panel) dataset for each zip code. The outcome variables represent counts of patients with any diagnoses within our target groups across all of their visits at the month/year/zip code level, divided by that zip code's pediatric population during that time period.

$$^{\ddagger}\theta_{downstream} = \theta_d = \frac{\theta_{ap} + \theta_{as}}{2}$$



$\theta_{\text{downstream}}$ is the angle alignment of wind with the downstream direction from the closest pollution point source to a school.

$\theta_{p\to s}$ is the angle indicating the direction of the school from the pollution point source with respect to due north.

θw_p and θw_s indicate the direction the wind is blowing with respect to due north at the pollution point source and at the school, respectively.

θ_{ap} and θ_{as} indicate the alignment of wind direction at the pollution point source and at the school, respectively, with $\theta p\to s$.

Fig 2. Diagram and explanation of terms used in computing $\theta_{\text{downstream}}$ (θ_d)

MODELING

We model each of our three health outcomes using an instrumental variables regression. This instrumental variables regression involves a two-stage process:

1. Model PM2.5 for every month and zip code using all of our fixed effects variables and include our instrumental variable derived from wind direction, $\theta_{\text{downstream}}{}^{\ddagger}$.

$$\widehat{PM}_{2.5} = \beta_0 + \beta_1 \theta_{downstream}{}^{\ddagger} + \beta_2 x_{year} + \beta_3 x_{month} + \beta_4 x_{county}$$

Eqn 1. Stage 1 regression equation

2. Separately model each of our three health outcomes using all of our fixed effects variables, and use the predicted PM2.5 as an additional explanatory variable.

$$Y^{\dagger} = \frac{\sum diagnoses^{\dagger}}{Pop_{zipcode,0-19}} = \delta_0 + \delta_1 \widehat{PM}_{2.5} + \delta_2 x_{year} + \delta_3 x_{month} + \delta_4 x_{county}$$

Eqn 2. Stage 2 regression equation

†DIAGNOSIS GROUPS

We aggregate diagnoses per patient and zip code/year/month combination:

- Group 1: **Hematopoietic Cancers**
- Group 2: **Type 1 Diabetes**
- Group 3: **Pediatric Vasculitis**
- Group 4: **Cardiorespiratory Disease**¶
- Group 5: **Injuries** (as control)¶
- Group 6: **None of the above**¶

RESULTS

Table 1 displays the results of the first stage regression:

coef_{θ_d}	p-value (θ_d)	95% CI
0.098	9.32e–13	[0.07, 0.12]

Table 1. Stage 1 regression results

The wind alignment's coefficient of 0.098 tells us that as θ_d increases by 1 (wind is more aligned with schools), the resulting PM2.5 prediction increases by 0.098. This small p-value tells us that there is an extremely small probability that the true impact of the wind alignment on PM2.5 is 0.

Next, we use the predicted PM2.5 amount from the first regression model and the same fixed effects to predict our three medical outcomes of interest. Because the PM2.5 value was predicted using the wind direction, which is uncorrelated with the outcomes of interest, the resulting coefficients (Table 2, below) provide a consistent estimate of the impact of PM2.5 on the health outcomes.

Disease Category	$\text{coef}_{\widehat{PM}_{2.5}}$	p-value ($\widehat{PM}_{2.5}$)	95% CI
Hematopoietic Cancer	0.133	0.061	[0.006, 0.273]
Type 1 Diabetes	0.041	3.0e–6	[0.024, 0.058]
Pediatric Vasculitis	–0.109	0.020	[–0.200, –0.017]
Cardioresp. Disease¶			
Injuries (control)¶			

Table 2. Stage 2 regression results

¶These diagnoses will be examined in the coming weeks.

Out of the outcomes we studied, PM2.5 has the largest impact on type-1 diabetes for patients under the age of 19 in the state of California. The coefficient of 0.041 tells us that as PM2.5 increases by 1 µg/m³, we expect type-1 diabetes hospital diagnoses to increase by 0.041 per month, per 1,000 children living in the zip code (p = 0.000003). The PM2.5 estimated impact on pediatric vasculitis suggests, surprisingly, that as PM2.5 increases by 1 µg/m³, pediatric vasculitis hospital visits would be expected to decrease by .109 visits per month per 1,000 children (p = 0.02; see caveat below). Finally, the estimated impact of PM2.5 on hematopoietic cancers was not statistically significant (p = 0.061) which means there is a 6% chance that PM2.5 has no impact on hematopoietic cancers.

Caveat: The results above are calculated using data to which noise has been added. This analysis will be run on a more complete dataset in the coming weeks.

FUTURE DIRECTION

Use a BEHRT model (a transformer model based on diagnosis history embeddings) to make predictions based on patient diagnosis history and residential environment.

Test different model structures in our instrumental regression, such as gradient-boosted trees.

Regress additional diagnosis groups on $\widehat{PM}_{2.5}$: cardiorespiratory diseases, to confirm an effect is seen; and injuries, to confirm that no effect is seen.

Citations: Al Brender JD, Maantay JA, Chakraborty J. Residential proximity to environmental hazards and adverse health outcomes. Am J Public Health. 2011;101 (Suppl 1): S37-S52. doi:10.2105/AJPH.2011.300183

Acknowledgements: Alberto Todeschini, PhD—UC Berkeley
Matthew Meyer, MD; Kyle Enfield, MD—UVA Health
Robert Davis, PhD—UVA Dept of Environmental Sciences

Contact: trevorj, mattslyons, anand.patel, michelleshen @berkeley.edu