

# Geographic and Socioeconomic Heterogeneity in the Benefits of Reducing Air Pollution in the United States

Tatyana Deryugina, Nolan Miller, David Molitor, Julian Reif<sup>1</sup>

May 2020

**Abstract:** Policies aimed at reducing the harmful effects of air pollution exposure typically focus on areas with high levels of pollution. However, if a population's vulnerability to air pollution is imperfectly correlated with current pollution levels, then this approach to air quality regulation may not efficiently target pollution reduction efforts. We examine the geographic and socioeconomic determinants of vulnerability to dying from acute exposure to fine particulate matter (PM<sub>2.5</sub>) pollution. We find that there is substantial local and regional variability in the share of individuals who are vulnerable to pollution both at the county and ZIP code level. Vulnerability tends to be negatively related to health and socioeconomic status. Surprisingly, we find that vulnerability is also negatively related to an area's average PM<sub>2.5</sub> pollution level, suggesting that basing air quality regulation only on current pollution levels may fail to effectively target regions with the most to gain by reducing exposure.

## I. Introduction

Recent research has found that acute pollution exposure is harmful to health even in areas where ambient pollution levels are generally low, such as in the US (e.g., Ward 2015; Knittel et al. 2016; Schlenker and Walker 2016; Deryugina et al. 2019). This research suggests that there may be substantial social benefits to further reductions in US air pollution. However, additional emissions reductions may require increasingly costly measures, making it crucial to understand where such reductions would be most beneficial.

---

<sup>1</sup> University of Illinois and NBER.

The benefits of air quality regulation in a region depend on many factors, including the amount by which air pollution is reduced, the vulnerability of the local population to air pollution, local population density, and, if the pollution-damage function is nonlinear, the initial level of pollution. Traditional approaches to air quality regulation have targeted regions that have high levels of pollution. For example, the Clean Air Act requires “non-attainment” areas that fail to meet air quality standards to take action to reduce pollution and to achieve attainment status as soon as possible, while areas that meet the standards do not need to take additional actions to further improve air quality. However, if pollution levels are imperfectly or negatively correlated with population vulnerability and density, areas with high pollution levels may not be the most cost-effective places to target for pollution reduction.

We investigate factors that predict elderly vulnerability to fine particulate matter (PM<sub>2.5</sub>) and measure how well they correlate with local PM<sub>2.5</sub> levels. By improving our understanding of the geographic and socioeconomic characteristics that matter for vulnerability, our results can help policymakers identify the most promising targets for air pollution reduction or for compliance and enforcement efforts.

Deryugina, Heutel, Miller, Molitor, and Reif (2019, henceforth DHMMR) show that there is substantial heterogeneity in vulnerability to acute PM<sub>2.5</sub> exposure in the US elderly population. While acute pollution exposure increases mortality among the elderly overall, a machine-learning-based analysis involving extensive individual and local characteristics estimates that acute PM<sub>2.5</sub> exposure increases the probability of death for only about 25 percent of Medicare beneficiaries.

In this paper, we extend DHMMR’s analysis to identify the geographic and socioeconomic correlates of such vulnerability and investigate the extent to which factors correlated with vulnerability are related to local pollution levels. If, for example, poor areas tend to attract pollution sources, such as factories and traffic, and poor people tend to be in worse health, then targeting pollution regulation at high-pollution areas may be an effective way of protecting individuals who are at the highest risk of pollution-related illness or death. However, if vulnerable populations do not tend to locate in high-pollution areas, then current pollution policy may be poorly targeted. As a result, existing pollution reduction efforts could be adapted to achieve greater increases in health, or similar increases in health could be achieved at lower resource costs.

Although the methods used in the DHMMR vulnerability prediction are computationally complex, the basic idea is straightforward. Following this approach, we generate two mortality

predictions for each elderly individual who was enrolled in Medicare in 2013. The first captures their likelihood of dying based on the experiences of similar people on days when they are exposed to high pollution, while the second predicts the likelihood of death based on the experiences of similar people on days when they are not exposed to high pollution. The average difference in the two predictions, which we refer to as the person's vulnerability index, represents the increased likelihood of the person dying on a day due to elevated acute PM<sub>2.5</sub> exposure.

After identifying those who are most and least vulnerable to death from pollution exposure, we compare the prevalence of various characteristics of individuals who are predicted to be highly vulnerable to PM<sub>2.5</sub> to individuals who are predicted to have low vulnerability to PM<sub>2.5</sub>. Importantly, we base our predictions on a large set of individual-level measures from Medicare data and ZIP-code-level socioeconomic factors from the US Census and related datasets. These rich data allow us to construct accurate and precise measures of vulnerability.

We find that the individuals identified in our data as most vulnerable to pollution are less healthy than the least vulnerable on a variety of measures, including the presence of chronic conditions such as Alzheimer's disease or related dementia, chronic obstructive pulmonary disease (COPD), lung cancer, chronic kidney disease, and congestive heart failure, as well as measures of health care use and spending. Geographically, we find that areas with high proportions of vulnerable individuals tend to form an L-shaped pattern, extending south from the Dakotas to Texas and then east along the Gulf Coast states. We also find significant heterogeneity at the county level within states and also at the ZIP code level within counties. Finally, we find that vulnerability and average PM<sub>2.5</sub> levels are negatively related despite the fact that average PM<sub>2.5</sub> levels are positively related to the prevalence of an array of adverse health conditions.

Overall, our analysis casts doubt on the presumption that a region's baseline pollution level is sufficient to target pollution reduction efforts—whether through regulation or direct expenditure—on those individuals and communities who will benefit the most, as baseline pollution and vulnerability to pollution are negatively related. In particular, regulations such as the Clean Air Act, which impose penalties on high-pollution areas but do not require reductions in average pollution or mitigation of pollution spikes in low-pollution areas, may fail to direct resources to their highest-benefit uses. Further, the substantial within-county heterogeneity in vulnerability that we identify suggests broad, geographically defined approaches are also likely to be imprecisely targeted and that additional attention should be paid to policies that account for

local population socioeconomic characteristics such as income, education, and health; local amenities such as hospital quality and capacity; and local environmental characteristics.

While DHMMR focused only on individuals living in counties with pollution monitors (902 counties in total), the sample we use in this paper includes beneficiaries who reside in any county in the conterminous US (3,101 counties in total), whether or not the county has a pollution monitor. Adding these new beneficiaries allows us to greatly expand our inquiry into geographic heterogeneity, as over two-thirds of US counties lack pollution monitors.

Our study is not without limitations. Most importantly, our analysis focuses on the elderly, a large vulnerable population likely to benefit from reductions in air pollution, and excludes working age adults and children. Prior studies have documented significant effects of air pollution on infant mortality, even in developed countries (Chay and Greenstone 2003; Knittel et al. 2016). While the elderly represent a large and important fraction of the US population, we emphasize that our results are not readily applied to younger age groups.

The rest of the paper is organized as follows. Section II provides a brief background on PM<sub>2.5</sub> and describes the data we use. Section III summarizes our analytical approach. Section IV presents the results, and Section V concludes.

## **II. Background and Data**

PM<sub>2.5</sub> is a mixture of various particles with diameters of less than 2.5 micrometers, including nitrates, sulfates, ammonium, and carbon (e.g., Kundu and Stone 2014). Manmade PM<sub>2.5</sub> comes from power plant and car emissions and can be carried for hundreds of miles away from where it is emitted. In many parts of the country, particularly the East, regional, rather than local, emissions make up a significant share of local particulate matter (Environmental Protection Agency 2004). The extent of pollution transport depends on a host of factors, including wind direction and speed, precipitation, and chemical reactions with other airborne molecules.

Many studies have examined the effect of air pollution on various health outcomes, including mortality. Much of the scientific literature focuses on the health effects of PM<sub>2.5</sub> because fine particulate matter can penetrate lung tissue and get into the bloodstream. Numerous epidemiological studies have documented a positive correlation between short-term exposure to particulate matter and mortality, especially from cardiovascular and respiratory disease (e.g., Pope and Dockery 2006; Samet et al. 2000; Environmental Protection Agency 2011). However, quasi-

experimental methodology such as the one we utilize, has been argued to be much more reliable than associational studies (Dominici et al. 2014).

Our health and health care use data come from Medicare administrative files. To inform our vulnerability index, we use the sample of all elderly beneficiaries aged 65 through 100 enrolled in Medicare in 2001–2013. We then focus our analysis on a single cohort: those aged 65–100 who were enrolled in Medicare in 2013. This sample comprises over 97 percent of elderly US residents that year.<sup>2</sup> Medicare enrollment files provide verified dates of death, age, sex, and county of residence. The Medicare Provider Analysis and Review (MedPAR) file provides data on health care use and cost for individuals enrolled in traditional fee-for-service Medicare. Detailed data on health care use are not available for individuals enrolled in Medicare Advantage managed care plans.<sup>3</sup> The MedPAR file, derived from Medicare Part A (facility) claims, provides information on each inpatient stay in a hospital or skilled nursing facility. The MedPAR data include information on the date of admission, length of stay, and total monetary cost of the stay.<sup>4</sup>

Spending on inpatient stays accounted for about 70 percent of all Medicare Part A costs and about 43 percent of all Medicare spending (including Parts A, B, and D) on elderly fee-for-service beneficiaries during 1999–2013, the years on which our machine-learning model is trained. We complement this dataset with data on outpatient ER visits that do not result in admission to the hospital from Medicare outpatient claims files, although we do not observe the cost of these visits. Because our unit of analysis is the county-day, we aggregate the Medicare data using patients' county of residence and the admission date (for inpatient stays) or the date of service (for outpatient ER visits).

The chronic conditions segment of the Master Beneficiary Summary File provides individual-level indicators for the presence of 27 different chronic conditions, such as heart disease, COPD, diabetes, and depression. These indicators are generated by professional medical coders who infer these conditions from a detailed examination of the claims data. Because they are based on claims data, this information is only available for beneficiaries in fee-for-service Medicare. And because it may take time for the relevant claims to appear in the data, the chronic

---

<sup>2</sup> Focusing on a single cohort ensures that each individual who was alive in 2013 appears in our data only once.

<sup>3</sup> In 2013, 28 percent of all Medicare beneficiaries were enrolled in Medicare Advantage plans.

<sup>4</sup> Our measure of cost is the total allowed charges due to the provider and includes all monetary costs of the stay, consisting of payments made by Medicare, the beneficiary, and/or another payer.

conditions indicators is most reliable for individuals who have been enrolled in fee-for-service Medicare for several years.

Air pollution levels are correlated with temperature and precipitation, which may have independent effects on mortality. Our analysis therefore controls for these variables to avoid confounding their effects with the effects of pollution. Our temperature and precipitation data come from Schlenker and Roberts (2009) and include total daily precipitation and maximum and minimum temperatures for each point on a 2.5-by-2.5-mile grid covering the conterminous US from 1999 to 2013. The Schlenker and Roberts (2009) data are derived from combining underlying data from PRISM and weather stations.<sup>5</sup> We aggregate to the county-day level by averaging the daily measures across all grid points in the county.

ZIP-code-level characteristics include various income and employment measures (e.g., median income, median home value, fraction of the population below the poverty line, labor force participation rate), measures of overall population health (e.g., the fraction of population with hearing or vision difficulties), and some other characteristics (e.g., travel time to work, prevalence of different heating fuels). These data are taken from the American Community Survey's 2007–2011 five-year estimates. Health-related variables such as disability and health coverage come from the 2008–2012 ACS, the first years that this information was included in the ACS.

Similarly, our county-level characteristics, which we correlate with the vulnerability index, come from a variety of sources, including Medicare administrative records, the US Census, the Behavioral Risk Factor Surveillance System (BRFSS), and variables used and constructed by Chetty and Hendren (2018) in their study of neighborhood impacts on intergenerational mobility. Characteristics we consider include average income; average Medicare spending; population health indicators such as the rate of smoking, exercise, and obesity; average temperatures; the crime rate; and measures of intergenerational income mobility. These characteristics are intended to capture an area's key environmental, economic, and public health conditions.

---

<sup>5</sup> See <http://www.prism.oregonstate.edu/> for the original PRISM dataset and <http://www.columbia.edu/~ws2162/links.html> for a detailed description of the daily data. Accessed February 26, 2020.

### III. Empirical Strategy

We are interested in estimating the causal effect of air pollution on mortality. Quantifying this effect in non-experimental data is complicated because air pollution is correlated with many other factors that matter for health. For example, traffic jams increase both pollution levels and stress, and low-income individuals are more likely to reside in high-pollution areas. Moreover, air pollution is not well measured: even monitored counties typically have only a few Environmental Protection Agency (EPA) pollution monitors. But an individual's air pollution exposure likely depends on finer measures of geography—such as which side of a highway she lives on. Such measurement errors can lead to biased estimates of the effect of air pollution even in settings where the variation in air pollution is as good as random.

To overcome these difficulties, our empirical strategy builds on DHMMR, who exploit quasi-random transport of PM<sub>2.5</sub> by the wind to estimate the mortality costs of acute air pollution exposure. Because changes in daily wind direction are unlikely to be related to other factors that affect health (such as traffic levels), this approach is likely to capture health effects attributable solely to air pollution. This approach also addresses measurement error concerns because it employs variation in air pollution that is blown in from far away and affects an entire area.

The amount of transported pollution is significant (Zhang et al. 2017). For example, the EPA estimates that most of the PM<sub>2.5</sub> in the Eastern United States was transported from hundreds of miles away (EPA 2004). DHMMR exploit this variation by instrumenting for daily PM<sub>2.5</sub> with the local wind direction. Their study shows that local wind direction is strongly predictive of changes in local PM<sub>2.5</sub>, even after conditioning on extensive controls for other atmospheric conditions and a host of fixed effects.

DHMMR's analysis focuses on areas with EPA monitors, which includes 902 counties covering about 70 percent of the national elderly population. Their instrumental variables design estimates that, on average, a one-unit increase in PM<sub>2.5</sub> (about 10 percent of the daily mean) increases same-day mortality by approximately 0.36 deaths per million beneficiaries in their sample. To estimate heterogeneity in the vulnerability to dying from acute air pollution exposure, they then apply a recently developed machine-learning method (Chernozhukov, Demirer, Duflo, and Fernandez-Val 2018, hereafter CDDF) to the rich set of characteristics available in the Medicare data. CDDF's method demonstrates how to estimate heterogeneous treatment effects using machine-learning techniques in the context of a binary treatment variable.

To form a binary treatment variable, DHMMR assign a person-day observation to the “treatment” group if the local wind direction on that day is associated with an above-median level of PM<sub>2.5</sub>, as measured in their first-stage specification. Otherwise, the observation is assigned to the “control” group. They then train a gradient-boosted decision tree algorithm (Chen and Guestrin 2016) to predict one-day mortality,  $Died_{it}$ , as a function of various measures of weather conditions, Census division fixed effects, local economic conditions, and individual-level characteristics, all denoted by  $Z_{it}$ . The model is estimated separately for observations in the treatment and control groups, resulting in two mortality prediction models.

In this paper, we apply the models estimated by DHMMR to the 2013 Medicare cohort (i.e., all individuals 65 and older who were alive and enrolled in Medicare at some point in 2013). Notably, our sample includes beneficiaries who reside in counties without pollution monitors and who were thus not included in DHMMR. The addition of these new beneficiaries allows us to greatly expand our inquiry into geographic heterogeneity, as about two-thirds of US counties lack pollution monitors. We can include these beneficiaries in our vulnerability analysis because, as we explain below, pollution data are not required to calculate the vulnerability index. (PM<sub>2.5</sub> data are required only to estimate the model, which has already been done in DHMMR.)

To ensure that we have reliable chronic condition indicators, we restrict our attention to beneficiaries who have been continuously enrolled in fee-for-service Medicare for at least two years. We create an observation for each day such an individual is alive and then predict their daily mortality probabilities using both the treatment group and control group prediction models.<sup>6</sup> The difference between these two predictions,  $\hat{S}_{it}(Z_{it})$ , represents the change in the observation’s predicted likelihood of death due to being exposed to a high-pollution wind direction, and it is referred to as a proxy predictor of the (true) conditional average treatment effect,  $s_0(Z_{it})$ . We then take the person-level average of these daily proxy predictors to calculate a single proxy predictor for each beneficiary,  $\bar{S}_i(Z_{it})$ .

$\hat{S}_{it}(Z_{it})$  and, by extension,  $\bar{S}_i(Z_{it})$  can be used to infer where in the distribution of treatment effects an observation lies. DHMMR estimate average treatment effects for various

---

<sup>6</sup> The mortality models from DHMMR include controls for two leads and two lags of the treatment indicator. Because we cannot determine treatment status for observations in counties without pollution monitors, we omit these controls from this paper. Omitting these high-level controls is unlikely to have any meaningful impact on our individual-level vulnerability index.



percentiles of the proxy predictor in their sample and conclude that about 25 percent of the Medicare population is vulnerable to acute fluctuations in PM<sub>2.5</sub>. Given this finding, we focus our analysis on individuals whose average proxy predictors place them in the top 25 percent of the overall distribution of  $\bar{S}_i(Z_{it})$  in the 2013 Medicare cohort. We hereafter refer to these individuals as “vulnerable to acute PM<sub>2.5</sub> exposure.”

## IV. Results

### IV.1 The geographic distribution of air pollution vulnerability

Table 1 compares the characteristics of 2013 elderly Medicare beneficiaries who are vulnerable to acute PM<sub>2.5</sub> exposure to the characteristics of those who are not vulnerable. On average, those who are vulnerable are almost four-and-a-half years older and are four percentage points more likely to be male. They are more than twice as likely to suffer from Alzheimer’s or related dementia, lung cancer, and congestive heart failure and are almost twice as likely to have chronic kidney disease or COPD. Consistent with their poor health indicators, beneficiaries who are vulnerable to acute PM<sub>2.5</sub> exposure also have substantially higher medical spending and are much more likely to have experienced various medical events, such as dialysis and hospice stays.

#### Insert Table 1 About Here

Figure 1 shows the geographic distribution of elderly who are vulnerable to acute PM<sub>2.5</sub> exposure as a percentage of the overall number of elderly Medicare beneficiaries in a given county.<sup>7</sup> Values below 25 percent indicate that a county’s beneficiaries are, on average, less vulnerable than the average fee-for-service beneficiary in the nation, while values above 25 percent indicate a disproportionately vulnerable population. As is readily apparent in Figure 1, there is a great deal of dispersion in this measure of vulnerability: some counties have less than 10 percent of their beneficiaries classified as vulnerable, while others have over 50 percent classified as vulnerable. In addition, there is substantial local variation in this measure, with some adjacent counties having very different scores. Although some of this variation may be due to noise, county-

---

<sup>7</sup> Recall that we define an individual as “vulnerable” if our model predicts that individual to be in the top 25 percent of the vulnerability distribution.

level variation can also be due to variation in factors like income and urbanity, which can vary discontinuously from one county to the next.

### **Insert Figure 1 About Here**

The counties shaded light blue or blue have between 20 and 30 percent of beneficiaries in the top 25 percent of vulnerability, which is near the 25 percent that would be expected if the county were representative of Medicare as a whole. Counties in lighter shades of blue and shades of purple represent counties where there are substantial deviations from the average of 25 percent. These deviations could be due to differences between the health or socioeconomic characteristics of the beneficiaries in that county and in Medicare overall, or they could be due to differences between characteristics of the county (e.g., healthcare infrastructure, baseline pollution, etc.) and those of the typical county where Medicare beneficiaries live. Our analysis below will identify some of these associations and illustrate the potential importance of both types of factors.

Figure 1 reveals several important patterns. First, areas with the highest concentration of vulnerable people tend to be concentrated in an L-shaped band running south from the Dakotas to Texas and then east through the Gulf Coast states. An additional group of counties with high concentrations of vulnerable people runs through eastern Kentucky and West Virginia. Second, there are marked differences in average vulnerability across states. For example, the West Coast states tend to have the lowest fraction of vulnerable people, while New England is in the middle, and Nebraska and West Virginia have high levels of vulnerability. Third, while many states, such as those along the Pacific Coast, are fairly uniform in terms of vulnerability, there are a number of states where there is significant within-state variation in vulnerability. For example, counties in western Kentucky tend to have lower concentrations of vulnerable people than eastern Kentucky, the Florida Panhandle has more vulnerability than the southern part of the state, and the Upper Peninsula of Michigan is more vulnerable than most of the Lower Peninsula.

Given the substantial amount of county-level heterogeneity in vulnerability depicted in Figure 1, a natural question is what drives this vulnerability. Because our analysis is not a causal one, we cannot speak directly to that question. But we can investigate the association between vulnerability and an array of county-level characteristics.

Figure 2 shows the relationship between the county-level share of vulnerable beneficiaries and various county-level characteristics. We emphasize that these relationships are descriptive and should not be interpreted causally. Some of these characteristics will be directly correlated with ZIP-code-level variables that were used in constructing the mortality models underlying our vulnerability index (e.g., median income and home value). Other characteristics are not used directly in constructing the vulnerability index (e.g., share urban, percent obese) but may nonetheless be correlated with characteristics that were included.

### **Insert Figure 2 About Here**

Each estimate reported in Figure 2 was obtained from a separate county-level regression where the outcome variable was the share of vulnerable beneficiaries in a county and the independent variable was a county characteristic. We consider one county characteristic at a time because some characteristics are highly correlated; in such cases, including multiple characteristics in the same regression can lead to a loss of precision and misleading conclusions. To make the results directly comparable to each other, we report coefficients and confidence intervals scaled by the interdecile range of a given characteristic (i.e., the difference between the 90th and 10th percentiles in our sample of counties). Thus, the results can be interpreted as the change in the share of the population that is vulnerable (in percentage points) when comparing a county in the 90th percentile of the distribution of a particular characteristic to a county in the 10th percentile of that distribution.

The associations shown in Figure 2 illustrate that vulnerability tends to be negatively related to health. Healthy behaviors, such as exercising, significantly decrease vulnerability, although in our elderly sample exercising may be acting primarily as an indicator of baseline health. In contrast, obesity and smoking prevalence are positively correlated with vulnerability, although this association may arise because obesity and smoking are correlated with other comorbidities rather than suggesting a causal effect on vulnerability. This possibility is supported by the fact that a high mortality rate is positively related to vulnerability. Indicators of high socioeconomic status are generally negatively related to vulnerability. High-income and high-median home values are both associated with low vulnerability, while a high poverty rate is positively related to vulnerability.

The climate variables, cooling degree days and heating degree days, are also related to vulnerability. Cooling degree days (which are high in generally hot places) are positively related to vulnerability, while heating degree days (which are high in generally cold places) are not significantly related to vulnerability. This result is somewhat surprising since studies have shown that both very hot and very cold days increase overall elderly mortality. The fact that heat interacts with vulnerability to air pollution in a way that cold does not suggests that different mechanisms likely underlie these two phenomena.

In addition to identifying areas with a high proportion of vulnerable people, optimal policy might also depend on identifying where the most vulnerable people live for a few reasons. First, while approximately 25 percent of the US elderly population are vulnerable to acute PM<sub>2.5</sub> exposure in the sense that they have a higher expected probability of death on polluted days than on clean days, for some of these people the increased risk is small, and thus it may be more effective to target air quality regulation at areas where the potential benefits of pollution reduction are large. Second, focusing on the top 25 percent of vulnerability may mask heterogeneity in the proportion of individuals who are most vulnerable to pollution exposure. To investigate these issues, we next turn to an analysis of geographic heterogeneity in this extremely vulnerable group.

Figure 3 shows the geographic distribution of beneficiaries who are extremely vulnerable to acute PM<sub>2.5</sub> exposure: those who are in the top 1 percent of the distribution of the proxy predictor  $\bar{S}(Z_{it})$ . The patterns are overall similar to those depicted in Figure 1, with areas with a high proportion of extremely vulnerable people falling along an L-shaped band from the Dakotas to Texas and then east along the Gulf Coast states.

### **Insert Figure 3 About Here**

Figure 1 and 3 illustrate the geographic distribution of vulnerable and “extremely” vulnerable populations, respectively. To quantify how correlated these two measures are, we estimate a population-weighted regression of the county-level share of vulnerable beneficiaries on the county-level share of extremely vulnerable beneficiaries. The R-squared from this regression is 0.61, indicating that considering the share of beneficiaries who are in the top 1 percent of the vulnerability index is highly but not perfectly informative about those who are in the top 25

percent.<sup>8</sup> An advantage of employing the broader (top 25 percent) definition of vulnerability is that it is subject to less measurement error in less-populated areas, so we focus on the top 25 percent for the remainder of the paper. There are, however, some differences in the patterns that we will briefly remark upon here. For example, at the southern tip of Texas, there are areas where the proportion of individuals in the top 1 percent of vulnerability is very high (purple) relative to other counties, but the proportion of individuals in the top 25 percent is more moderate. The opposite is also true, with counties where the relative frequency of individuals in the top 1 percent is low but the frequency of individuals in the top 25 percent is moderate.

We next turn to a ZIP-code-level analysis to illustrate heterogeneity at a very granular level. Figure 4 shows the geographic distribution of vulnerable beneficiaries (those in the top 25 percent of the  $\bar{S}(Z_{it})$  distribution) at the ZIP code level. To comply with disclosure rules, we do not show ZIP codes that have fewer than 100 Medicare beneficiaries in 2013. We observe 32,331 ZIP codes in our data, 21,506 of which have at least 100 Medicare beneficiaries.

#### **Insert Figure 4 About here**

Comparing the ZIP-code-level map in Figure 4 to the county-level map in Figure 1 reveals the existence of significant within-county heterogeneity. For example, the northern part of the lower peninsula of Michigan appears to have low vulnerability in Figure 1's county-level map, but Figure 4 reveals several highly vulnerable ZIP codes. Similarly, seen at the county level, Maine appears to be uniformly moderately vulnerable, while the ZIP-code-level analysis reveals a mix of high- (purple) and low- (light blue) vulnerability ZIP codes. While some of this variation may be due to noise, many of the important correlates of vulnerability identified in Figure 2, such as income, are known to vary within counties.

To illustrate the degree of ZIP-code-level variability in the share of vulnerable beneficiaries, Appendix Figures A.3–A.7 display ZIP-code-level maps of five commuting zones,

---

<sup>8</sup> Analogous to Figure 2, Figure A.1 shows the correlation between various county-level characteristics and the share of beneficiaries in the top 1 percent of vulnerability. Figure A.2 presents a scatterplot between the county-level share of beneficiaries in the top 25 percent of vulnerability and the share of beneficiaries in the top 1 percent of vulnerability.

some of which exhibit a lot of variability and some of which have very little variability.<sup>9</sup> To quantify the amount of within- versus across-county variation more systematically, we regress the ZIP-code-level share of vulnerable beneficiaries (including ZIP codes with less than 100 beneficiaries) on county fixed effects, weighting by the number of beneficiaries in that ZIP code. The R-squared in this regression is 0.33, suggesting that the majority (67 percent) of the ZIP-code-level variation depends on within-county differences. We perform a similar exercise using the share of beneficiaries who are extremely vulnerable (top 1 percent) and find that 76 percent of the ZIP-code-level variation depends on within-county differences.

## IV.2 Vulnerability and pollution levels

Given the high degree of geographic variation in vulnerability, it is natural to investigate whether that variation is correlated with variation in underlying pollution levels. For example, it may be that a given pollution shock is more deadly in regions that already have high pollution levels. Indeed, this hypothesis is at least consistent with the idea that pollution regulation should be targeted at locations with high pollution levels.

To give a sense of the geographic distribution of pollution, the map in Figure 5 shows the average annual PM<sub>2.5</sub> level for each US county in 2013. PM<sub>2.5</sub> pollution tends to be highest in California and in the Rust Belt states ranging from Illinois to Pennsylvania and lowest along the Rocky Mountains.

### Insert Figure 5 About Here

Figure 6 shows the county-level relationship between the share of elderly beneficiaries who are vulnerable and 2013 PM<sub>2.5</sub> levels, with a population-weighted trend line drawn to aid in visualizing the statistical relationship between the two. Perhaps surprisingly, less polluted counties tend to have *more* vulnerable beneficiaries: for each one-unit increase in average PM<sub>2.5</sub> levels, the share of vulnerable beneficiaries decreases by 0.83 percentage points. While this correlation could be coincidental (e.g., urban areas may be more polluted and attract more frail elderly because of superior medical care), another potential explanation is that those who are vulnerable to air

---

<sup>9</sup> Commuting Zones (CZs) are geographies similar to Metropolitan Statistical Areas (MSAs) that group nearby areas based on commuting patterns. For our purposes, CZs are preferred due to their superior coverage of rural areas and the fact that they are defined at the supra-county level.

pollution explicitly avoid more polluted areas. Indeed, residential sorting on the basis of air pollution levels has been documented in numerous prior studies (see Banzhaf et al. 2019 for a review). Regardless of the mechanism, the negative relationship lends additional support to the idea that targeting regulation at highly polluted areas may be less beneficial for population health than simple intuition would suggest. In addition, if we were to interpret this relationship causally, it suggests that reducing average pollution levels makes individuals more vulnerable to pollution spikes and vice versa.

### **Insert Figure 6 About Here**

As shown in Table 1, the Medicare beneficiaries who are most vulnerable to air pollution are less healthy than the average beneficiary. If beneficiaries in good general health are more likely to reside in areas with high levels of air pollution, then this might explain why we find an inverse relationship between a county's pollution levels and its average vulnerability.

We investigate that possibility by estimating the correlations between PM<sub>2.5</sub> levels and the other predictors of vulnerability we identified in Table 1. Figure 7 displays scatterplots of annual PM<sub>2.5</sub> levels against the county-level share of beneficiaries with a particular chronic condition. Although overall vulnerability is negatively correlated with PM<sub>2.5</sub> levels, more polluted counties, on average, have a higher share of beneficiaries with congestive heart failure (panel (a)), stroke (panel (b)), and Alzheimer's/dementia (panel (e)). More polluted counties also have higher average Medicare spending (panel (f)). However, we do not detect a significant relationship between ambient pollution levels and COPD (panel (c)) or lung cancer (panel (d)). Overall, these results suggest that the negative relationship between ambient PM<sub>2.5</sub> and vulnerability is not driven by chronic condition or average total Medicare spending.

### **Insert Figure 7 About Here**

The lack of a significant relationship between background pollution levels and COPD is particularly interesting since the harmful effects of the small particulates comprising PM<sub>2.5</sub> are thought to arise when the particles are inhaled and irritate the lungs. The fact that we do not find increased incidence of COPD in areas with high pollution levels is consequently surprising. This

null result could be due to the presence of confounders that are correlated with PM<sub>2.5</sub> levels. Alternatively, the impact of high pollution levels on the lungs could manifest in some way that is not classified as COPD. In either case, this issue warrants further study.

## **V. Conclusion**

This paper has explored the socioeconomic and geographic correlates of vulnerability to acute PM<sub>2.5</sub> exposure in the United States. Building on the analysis in DHMMR, we apply the model from that paper to the 2013 Medicare cohort. While DHMMR was restricted to 902 counties containing pollution monitors, our sample includes all Medicare beneficiaries living in the conterminous US (3,101 counties in total), which permits a detailed investigation of geographic heterogeneity. Our paper computes a proxy indicator for the conditional average treatment effect for each individual in our data and uses that to classify individuals as vulnerable or not vulnerable to acute air pollution.

As one might expect, we find that vulnerability is positively and significantly associated with a range of health indicators. Individuals in the top quartile of vulnerability are older, more likely to be male, and more likely to exhibit chronic conditions such as Alzheimer's disease or related dementia, chronic kidney disease, COPD, congestive heart failure, or lung cancer. Highly vulnerable individuals are also likely to spend more on health care and to consume more health care services.

We aggregate across individuals within a particular geographic area to investigate geographic heterogeneity. At the county level, we find a large degree of variation in the share of individuals in the top quartile of vulnerability, ranging from below 5 percent to above 50 percent. The areas with the highest proportion of individuals in the highly vulnerable category lie in an L-shaped band that ranges from the Dakotas south to Texas and then eastward through the Gulf Coast states toward Georgia and Northern Florida. An additional group of highly vulnerable areas fall in eastern Kentucky and West Virginia. In contrast, many of the counties in New England and Pacific Coast states have lower-than-expected shares of vulnerable residents.

Given the large amount of county-level heterogeneity, we next turn toward investigating the geographic and socioeconomic correlates. As might be expected from the individual-level analysis, we find that vulnerability and health tend to be positively correlated at the county level as well. A high share of individuals who report exercising reduces the share of vulnerable



individuals in the county, while smoking, obesity, and advanced age all tend to increase it, as does having a high local elderly mortality rate. The relationship between health care infrastructure and vulnerability tends to be more mixed, with high numbers of physicians per capita and high hospital quality correlating with low vulnerability, while having a high amount of hospital beds per capita and high Medicare spending per beneficiary are both correlated with higher vulnerability. The reasons for this discrepancy are not obvious, although reverse causation likely plays a role: areas with more vulnerable people will tend to have higher Medicare spending and higher mortality.

Turning to socioeconomic indicators, high income and high home values correspond to low vulnerability, while high poverty corresponds to high vulnerability. Having a large population and a high proportion of individuals living in urban areas are associated with lower vulnerability. Interestingly, areas with high levels of government services, as measured by local government spending per capita and local taxation per capita, tend to have lower levels of vulnerability.

Somewhat surprisingly, we find that the share of vulnerable individuals within a county is negatively related to baseline pollution, although it is positively related to various measures of poor health. Although the exact mechanism underlying this pattern is beyond the scope of this paper, it suggests that using high pollution as a basis for targeting air pollution efforts, as is done under the Clean Air Acts and other environmental regulations, may lead to misallocation of resources.

Although our study sheds substantial light on the geographic and socioeconomic heterogeneity in vulnerability to air pollution, some caveats are in order. First, as we have stated throughout the paper, our analysis is not causal. Nevertheless, the correlational patterns we identify in the paper may provide inspiration for future causal investigations. Second, our study is limited to the elderly. Although there is substantial evidence that the elderly are particularly vulnerable to pollution shocks, pollution has also been shown to increase infant mortality. To the extent that patterns of infant mortality differ from those of the elderly, these differences should also be taken into account by policymakers seeking to direct resources toward pollution reduction. Finally, although our analysis is based on a large sample of elderly Medicare beneficiaries from across the United States, our vulnerability computations are based only on a single year (2013). If vulnerability changes over time, then our 2013 analysis may not fully characterize current (or future) vulnerability.

## REFERENCES

- Banzhaf, Spencer, Lala Ma, and Christopher Timmins.** 2019. “Environmental Justice: The Economics of Race, Place, and Pollution.” *Journal of Economic Perspectives*, 33(1): 185–208.
- Chay, Kenneth, and Michael Greenstone.** 2003. “The Impact of Air Pollution on Infant Mortality: Evidence from Geographic Variation in Pollution Shocks Induced by a Recession.” *The Quarterly Journal of Economics*, 118(3): 1121–1167.
- 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.
- Chetty, Raj, and Nathaniel Hendren.** 2018. “The Impacts of Neighborhoods on Intergenerational Mobility II: County-Level Estimates.” *The Quarterly Journal of Economics*, 133(3): 1163–1228.
- Deryugina, Tatyana, Garth Heutel, Nolan Miller, David Molitor, and Julian Reif.** 2019. “The Mortality and Medical Costs of Air Pollution: Evidence from Changes in Wind Direction.” *American Economic Review*, 109(12): 4178–4219.
- Dominici, Francesca, Michael Greenstone, and Cass Sunstein.** 2014. “Particulate Matter Matters.” *Science*, 344(6181): 257–259.
- Environmental Protection Agency.** 2004. “The Particle Pollution Report. Current Understanding of Air Quality and Emissions through 2003.” US Environmental Protection Agency, Washington, DC.
- Environmental Protection Agency.** 2011. “The Benefits and Costs of the Clean Air Act from 1990–2020.” US Environmental Protection Agency, Washington, DC.
- Knittel, Christopher, Douglas Miller, and Nicholas Sanders.** 2016. “Caution, Drivers! Children Present: Traffic, Pollution, and Infant Health.” *Review of Economics and Statistics* 98(2): 350–366.
- Kundu, Shuvashish, and Elizabeth A. Stone.** 2014. “Composition and Sources of Fine Particulate Matter Across Urban and Rural Sites in the Midwestern United States.” *Environmental Science: Processes and Impacts*, 16(6): 1360–1370.
- Pope, C. Arden, and Douglas Dockery.** 2006. “Health Effects of Fine Particulate Air Pollution: Lines That Connect.” *Journal of the Air and Waste Management Association*, 56(6): 709–742.
- Samet, Jonathan, Francesca Dominici, Frank Curriero, Ivan Coursac, and Scott Zeger.** 2000. “Fine Particulate Air Pollution and Mortality in 20 US Cities, 1987–1994.” *New England Journal of Medicine*, 343(24): 1742–1749.
- Schlenker, Wolfram, and Michael J. Roberts.** 2009. “Nonlinear Temperature Effects Indicate Severe Damages to US Crop Yields under Climate Change.” *Proceedings of the National Academy of Sciences of the United States of America*, 106(37): 15594–15598.
- Schlenker, Wolfram, and W. Reed Walker.** 2016. “Airports, Air Pollution, and Contemporaneous Health.” *Review of Economic Studies*, 83(2): 768–809.

**Ward, Courtney.** 2015. “It’s an Ill Wind: The Effect of Fine Particulate Air Pollution on Respiratory Hospitalizations.” *Canadian Journal of Economics*, 48(5): 1694–1732.

**Zhang, Qiang, Xujia Jiang, Dan Tong, Steven Davis, Hongyan Zhao, Guannan Geng, Tong Feng, et al.** 2017. “Transboundary Health Impacts of Transported Global Air Pollution and International Trade.” *Nature*, 543(7647): 705–209.

## FIGURES AND TABLES

Table 1: Summary statistics for the Medicare beneficiaries most and least affected by pollution in 2013

Outcome	(1) Bottom 75%	(2) Top 25%	(3) Difference
<b>Demographics</b>			
Age (years)	75.3	79.7	4.42*** (0.00435)
Male	0.421	0.461	0.0408*** (0.000294)
<b>Chronic conditions</b>			
Alzheimer's or dementia	0.0946	0.239	0.144*** (0.000197)
Chronic kidney disease	0.176	0.343	0.166*** (0.000241)
COPD	0.197	0.386	0.189*** (0.00025)
Heart failure	0.192	0.412	0.219*** (0.00025)
Lung cancer	0.00965	0.0251	0.0155*** (0.0000685)
<b>Medical spending (dollars)</b>			
Durable medical equipment	160	333	173*** (0.438)
Hospice	148	394	245*** (1.69)
Hospital outpatient	1,177	2,246	1,068*** (2.63)
Part B drug	277	637	360*** (3.11)
Part B other	118	264	146*** (0.710)
<b>Medical events</b>			
Dialysis	0.0522	0.153	0.101*** (0.00073)
Durable medical equipment	2.17	4.27	2.10*** (0.00399)
Hospice stays	0.00626	0.0163	0.010*** (0.0000553)
Part B drug	2.54	3.90	1.36*** (0.00404)
Part B evaluation and management	4.25	8.86	4.61*** (0.00828)

Notes: Column (1) presents means for person-day observations predicted to have a below-median treatment effect. Column (2) presents means for those in the top 25 percent. Column (3) reports the difference between columns (2) and (1). Medical spending and medical events are measured over the calendar year prior to the date of the observation. Hospice stays are defined as the number of unique admissions. For all other medical events, the event is defined as each line item on the insurance claim that contains the relevant service. COPD stands for chronic obstructive pulmonary disease. Standard errors, clustered by county, are reported in parentheses. \*\*\*  $p < 0.01$ , \*\*  $p < 0.05$ , \*  $p < 0.10$ .

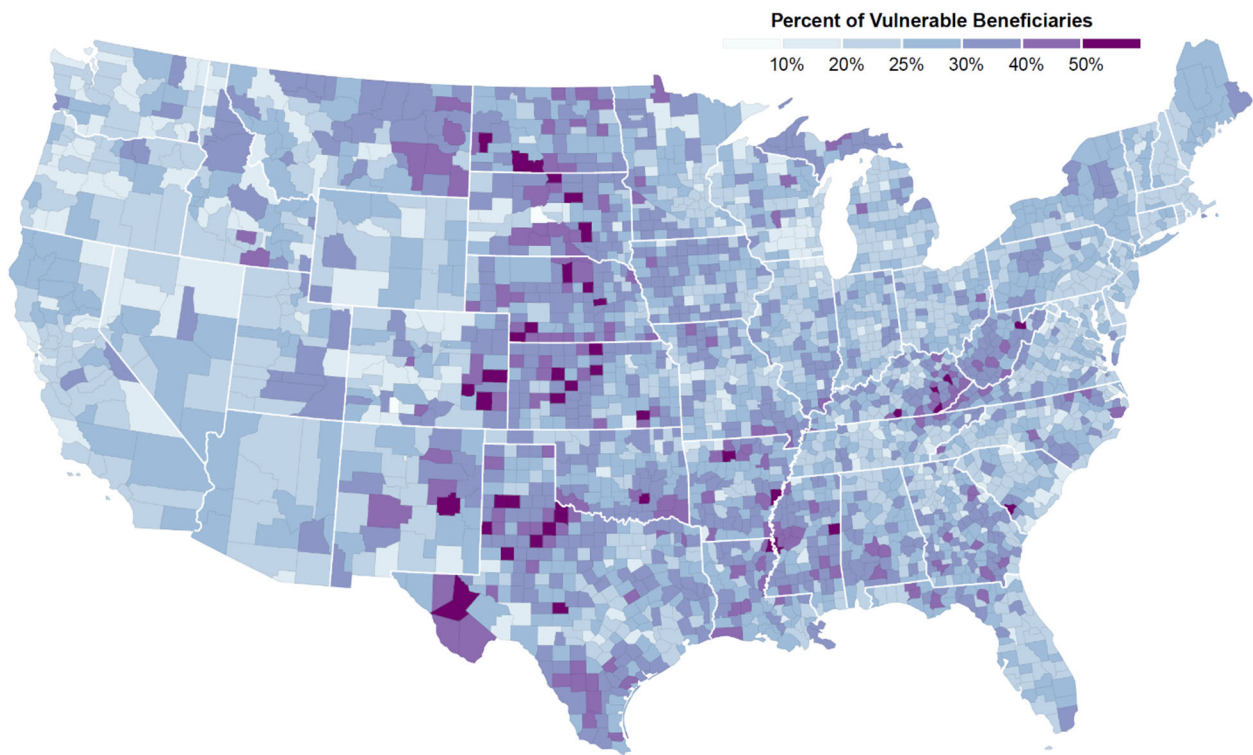


Figure 1. The map shows the fraction of Medicare beneficiaries in each county who were vulnerable to acute PM<sub>2.5</sub> exposure (i.e., were in the top 25 percent of the acute PM<sub>2.5</sub> vulnerability index) in 2013.

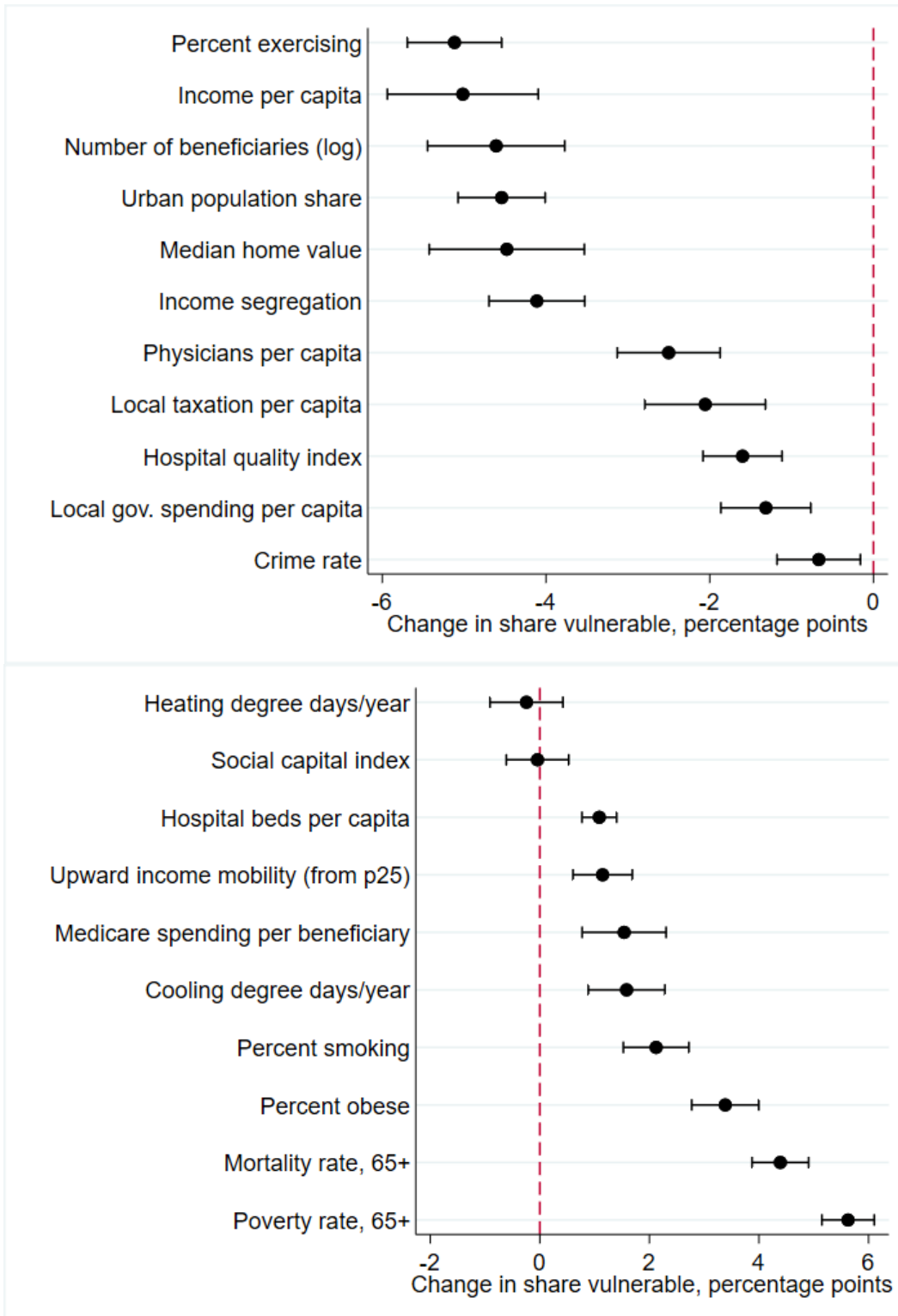


Figure 2. The figure shows correlations between the share of beneficiaries who were vulnerable to acute PM<sub>2.5</sub> exposure (i.e., in the top 25 percent of the acute PM<sub>2.5</sub> vulnerability index) in 2013 and county-level characteristics. Each estimate is from a separate county-level regression of the share vulnerable on the given characteristic.

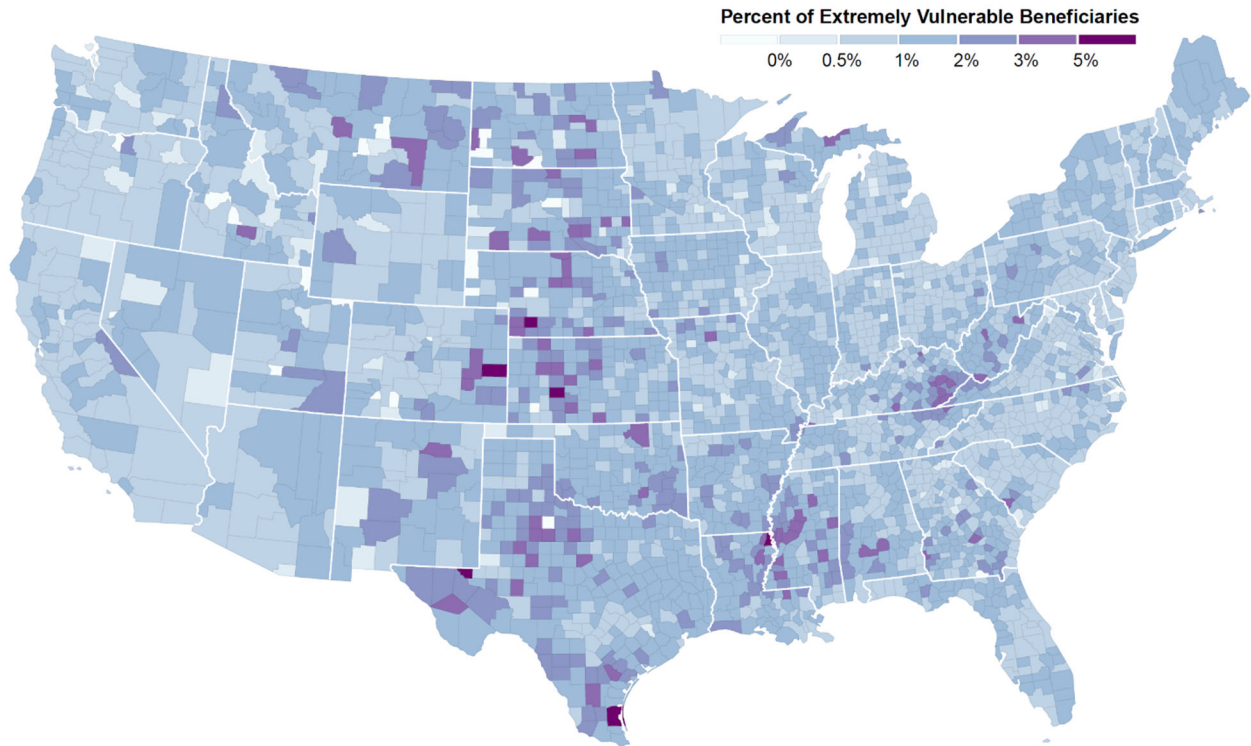


Figure 3. The map shows the fraction of Medicare beneficiaries in each county who were extremely vulnerable to acute PM<sub>2.5</sub> exposure (i.e., were in the top 1 percent of the acute PM<sub>2.5</sub> vulnerability index) in 2013.



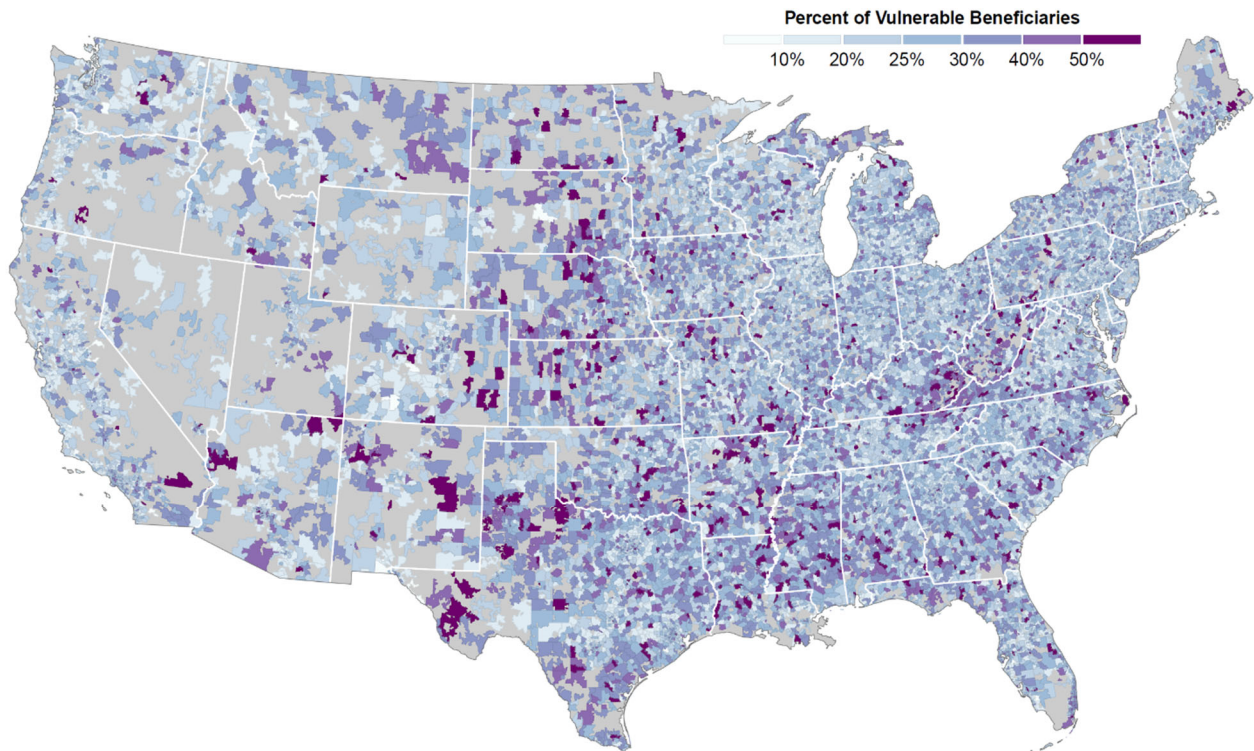


Figure 4. The map shows the fraction of Medicare beneficiaries in each ZIP code (ZCTA) who were vulnerable to acute PM<sub>2.5</sub> exposure (i.e., in the top 25 percent of the acute PM<sub>2.5</sub> vulnerability index) in 2013. Grey areas indicate regions not in a ZCTA or with fewer than 100 beneficiaries.



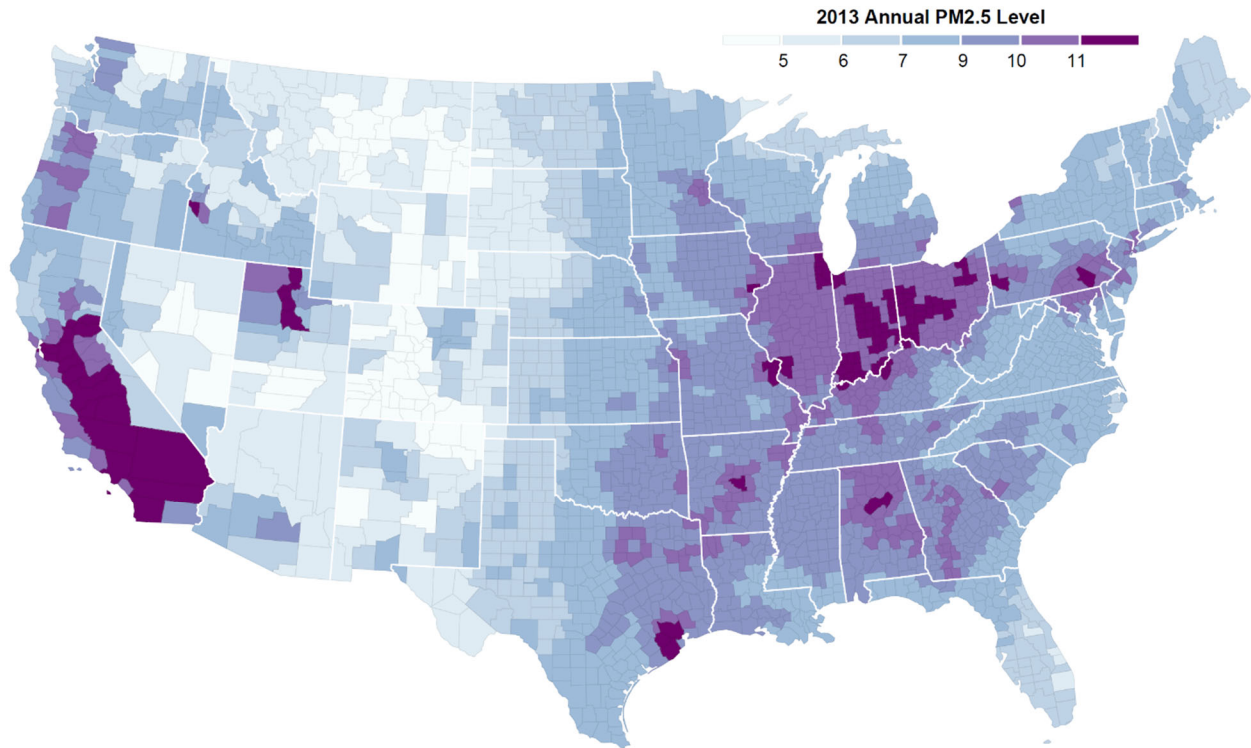


Figure 5. The map shows county-level annual PM<sub>2.5</sub> in 2013. The PM<sub>2.5</sub> measure is provided by the CDC's National Environmental Public Health Tracking Network and was created using monitor data when available and modeled estimates for days or counties that do not have monitor data.

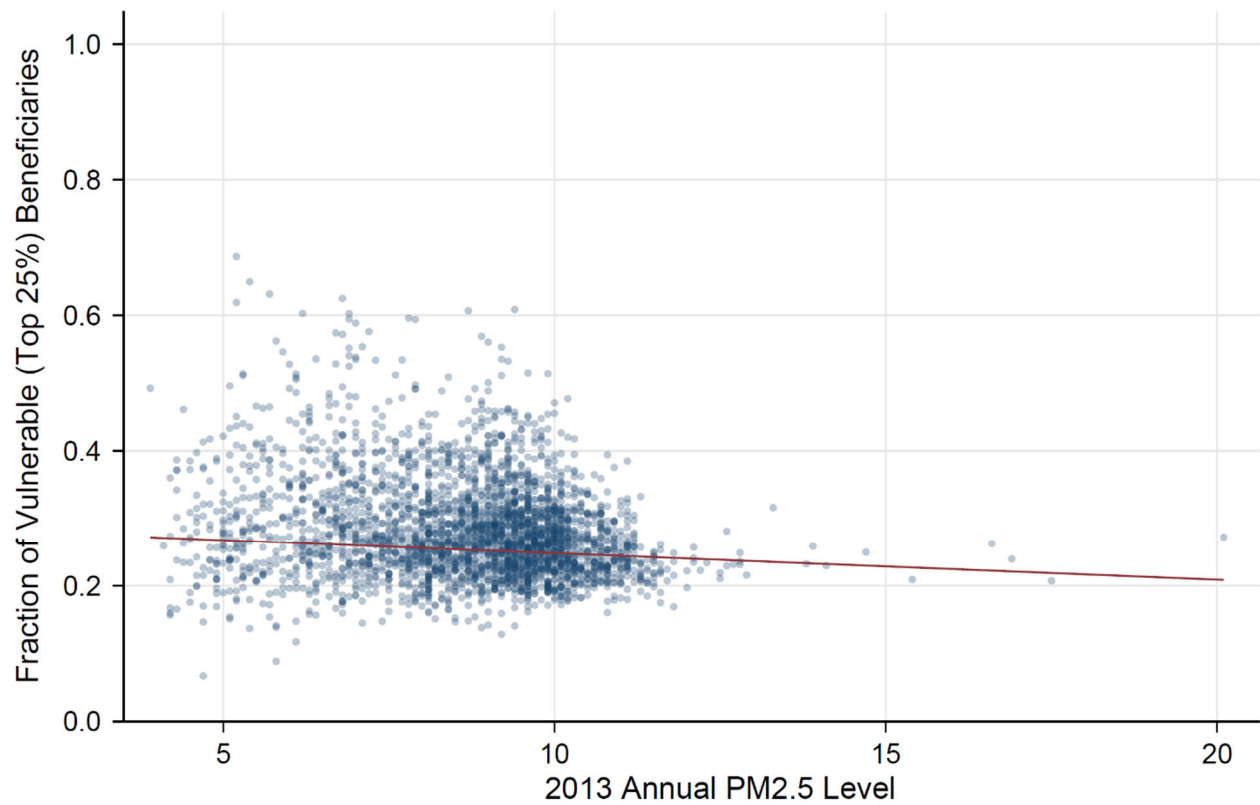
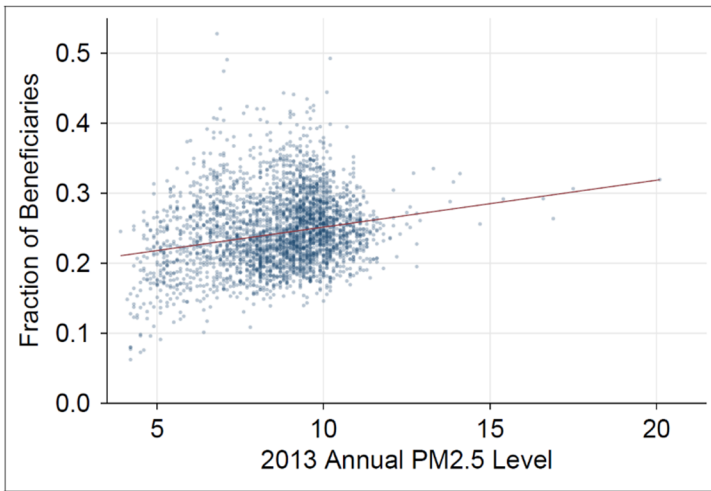
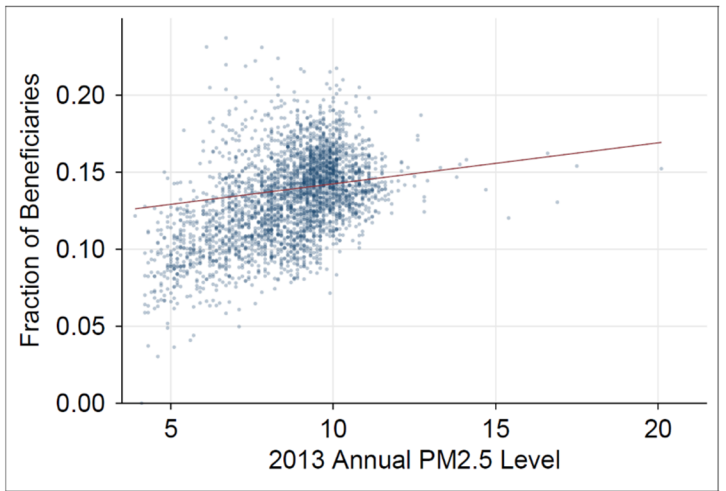


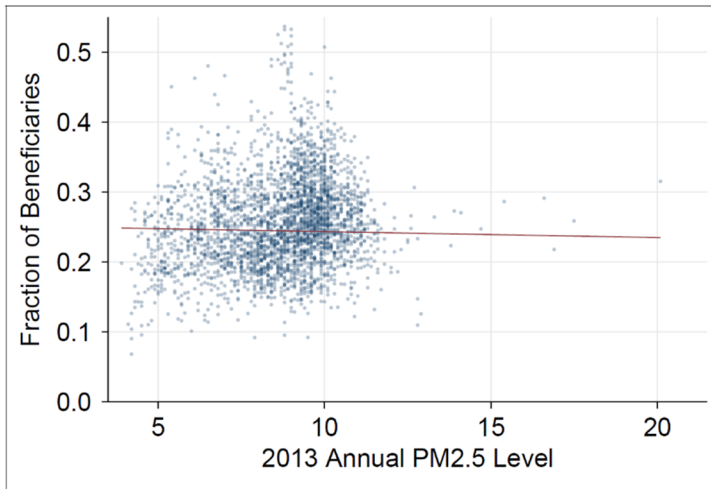
Figure 6. Each dot indicates the annual average ambient concentration of PM<sub>2.5</sub> in micrograms per cubic meter ( $\mu\text{g}/\text{m}^3$ ) and the fraction of Medicare beneficiaries who were vulnerable to acute PM<sub>2.5</sub> exposure for a county, in 2013.



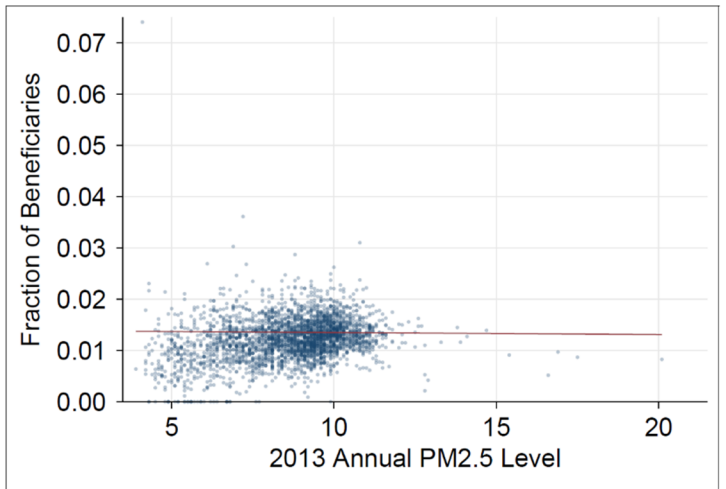
(a) Heart Failure



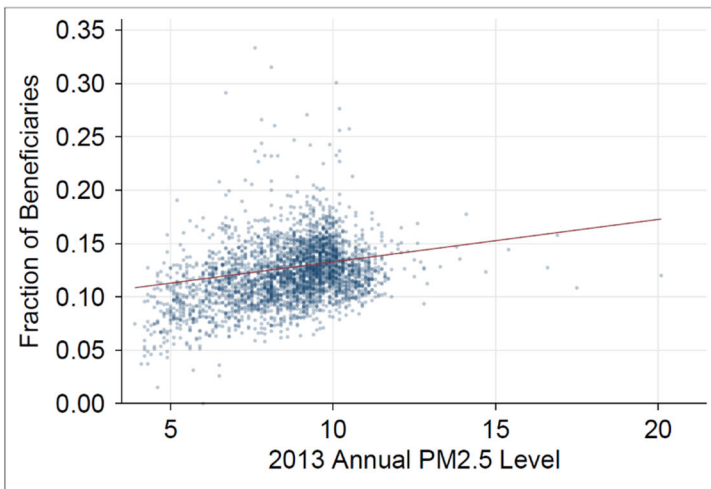
(b) Stroke



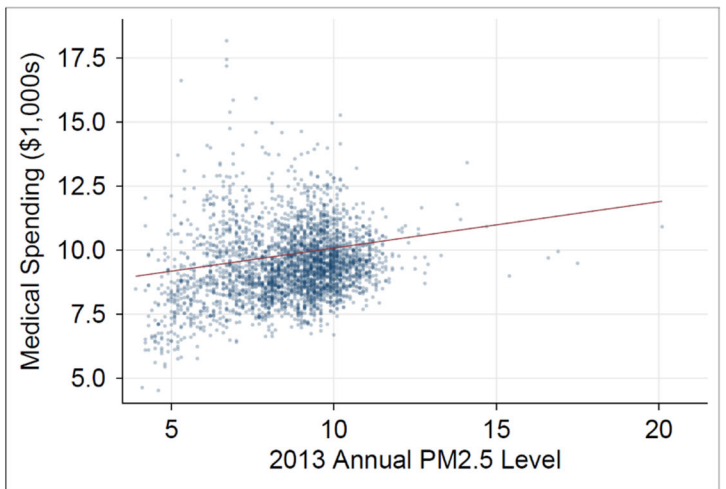
(c) COPD



(d) Lung Cancer



(e) Alzheimer's/Dementia



(f) Medical Spending, Prior Year

Figure 7. Each dot indicates the annual average ambient concentration of PM<sub>2.5</sub> in micrograms per cubic meter ( $\mu\text{g}/\text{m}^3$ ) and the fraction of Medicare beneficiaries with certain chronic conditions or average Medicare spending in a county, in 2013.

## APPENDIX

Table A.1: Summary of county-level characteristics

	10th percentile	Median	Mean	90th percentile
Heating degree days/year	1459	4600	4361	6904
Cooling degree days/year	471.1	1181	1421	2836
Hospital quality index	0.709	0.782	0.779	0.856
Hospital beds per capita	0.930	2.925	3.406	5.938
Physicians per capita	0.576	1.965	2.349	4.474
Urban population share	0.229	0.741	0.675	0.992
Poverty rate, 65+	5.577	8.753	10.11	17.04
Percent exercising	67.05	74.34	74.30	81.49
Percent obese	15.56	20.32	20.78	26.60
Percent smoking	16.86	21.68	21.94	27.04
Crime rate	2.801	6.786	6.981	11.23
Social capital index	-1.714	-0.350	-0.315	1.060
Local gov. spending per capita (1000s)	1.364	2.191	2.320	3.375
Local taxation per capita (1000s)	0.346	0.727	0.795	1.347
Income segregation	0.006	0.046	0.051	0.100
Upward income mobility (from p25)	-0.528	0.029	0.033	0.582
Median home value (1000s)	58.70	95.80	113.2	195.8
Income per capita (1000s)	15.18	19.79	20.73	27.91
Number of beneficiaries (log)	7.639	9.124	9.215	10.84
Medicare spending per beneficiary	9.219	10.91	11.16	13.39
Mortality rate, 65+	0.049	0.053	0.053	0.058

Notes: The table shows the population-weighted 10th percentiles, medians, means, and 90th percentiles of county-level characteristics used in the analysis.

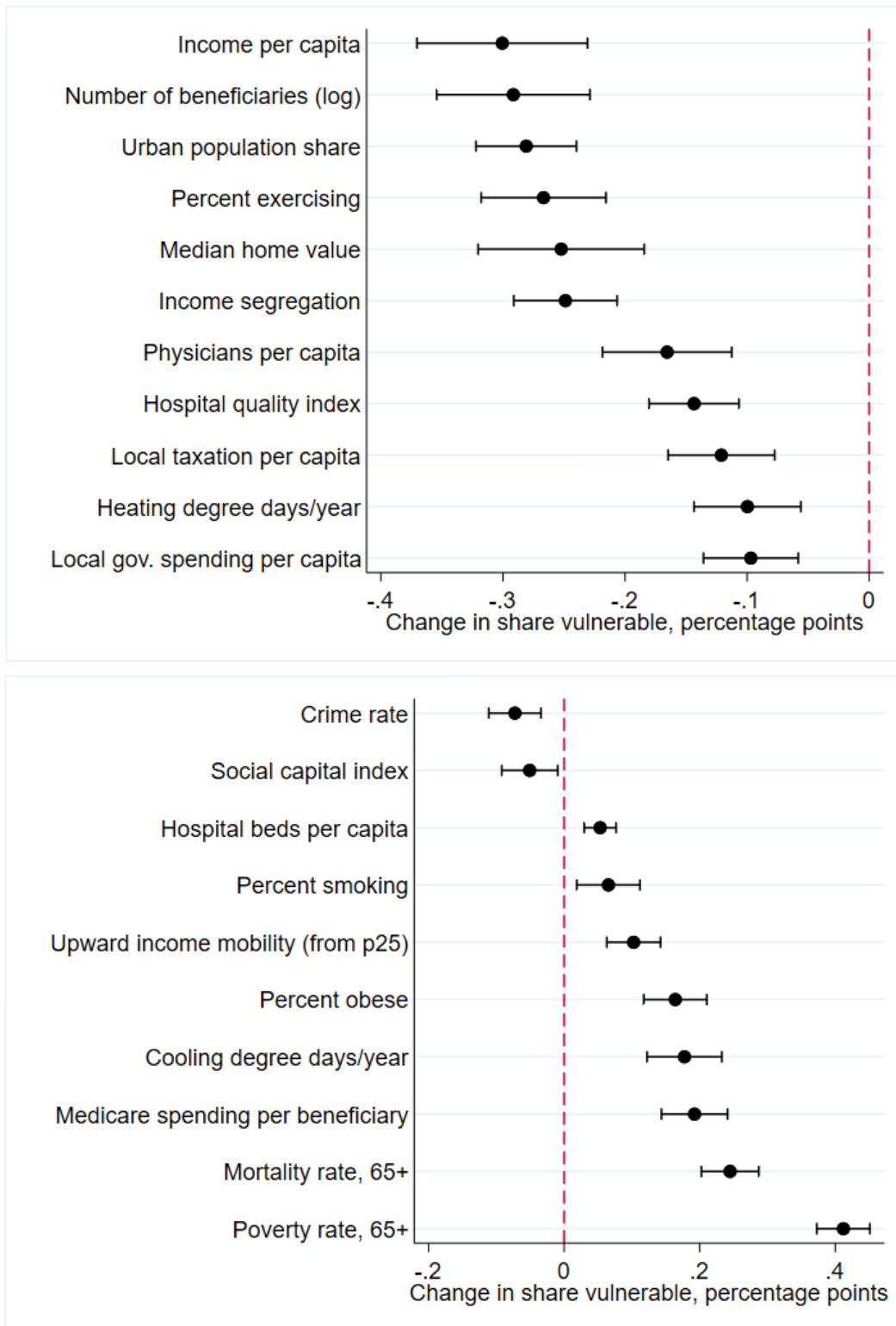


Figure A.1. The figure shows correlations between the share of beneficiaries who were extremely vulnerable to acute PM<sub>2.5</sub> exposure (i.e., in the top 1 percent of the acute PM<sub>2.5</sub> vulnerability index) in 2013 and county-level characteristics. Each estimate is from a separate county-level regression of the share extremely vulnerable on the given characteristic.

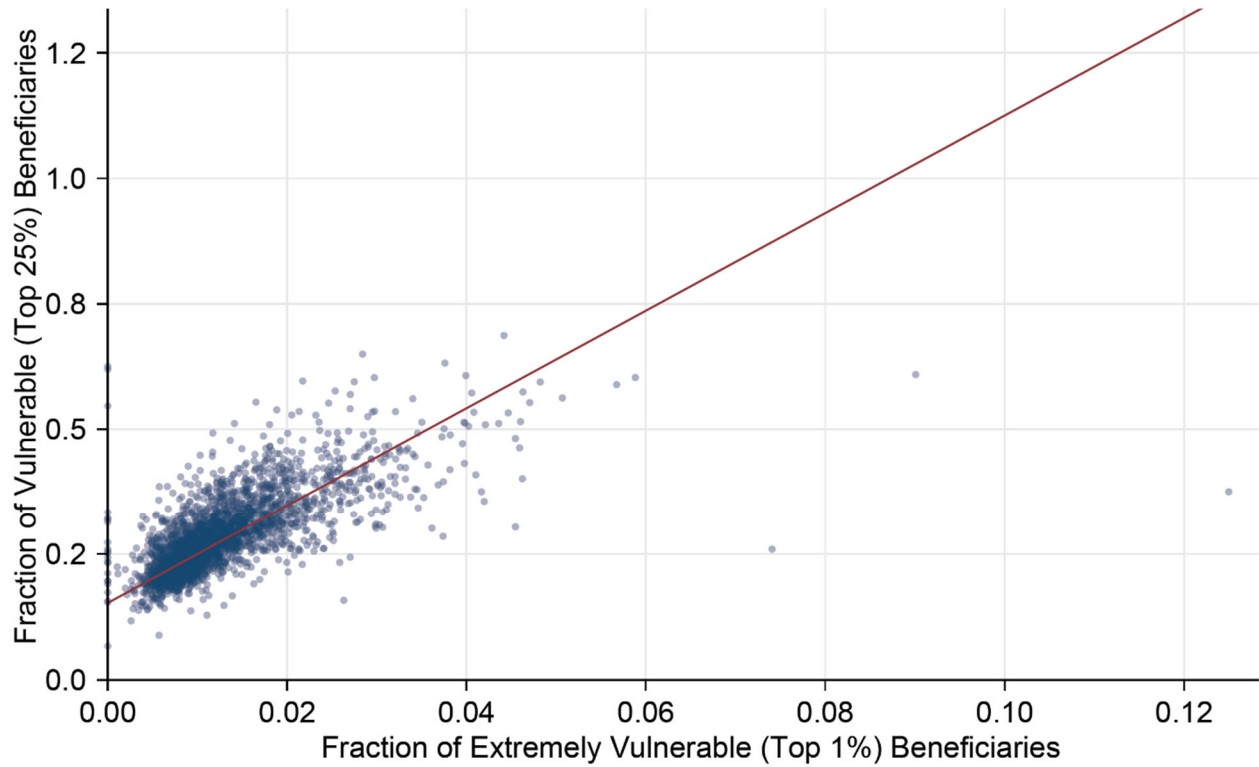


Figure A.2. Each dot represents a county and indicates the fraction of Medicare beneficiaries who were in the top 1 percent of the acute PM<sub>2.5</sub> vulnerability index (“extremely vulnerable”) and the fraction of beneficiaries who were in the top 25 percent of the vulnerability index (“vulnerable”) in 2013.

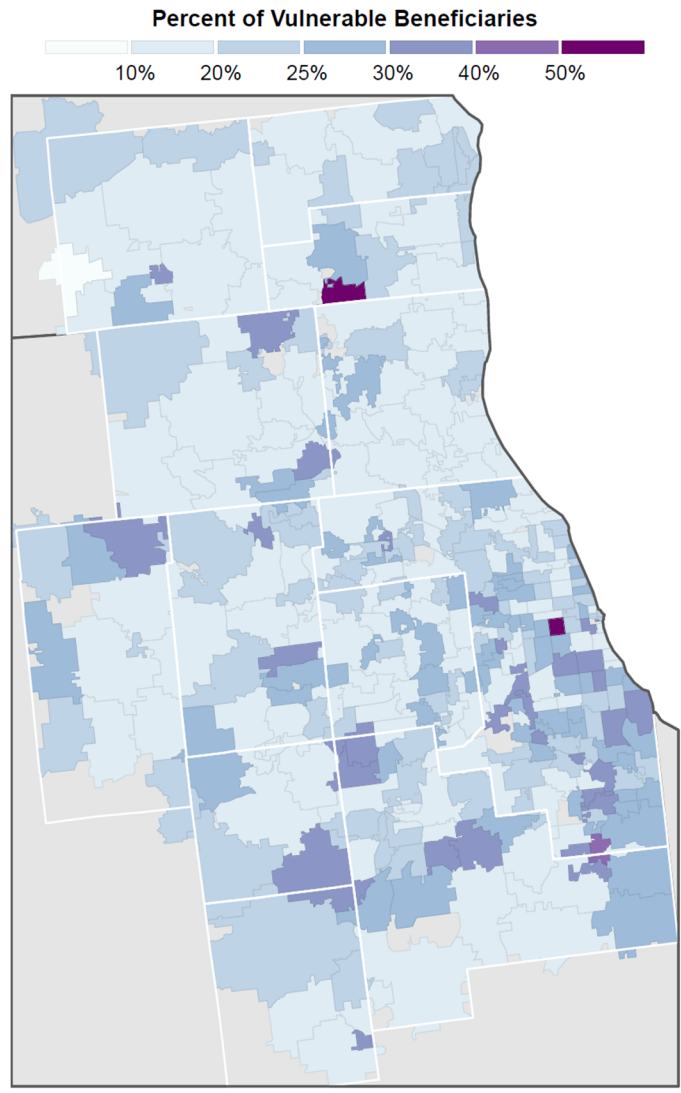


Figure A.3. The map reports ZIP-code-level vulnerability to acute PM<sub>2.5</sub> exposure for all counties in the Commuting Zone containing the Chicago-Naperville-Joliet, IL Metropolitan Division. ZIP code shading indicates the fraction of Medicare beneficiaries in each ZIP code who were vulnerable to acute PM<sub>2.5</sub> exposure (i.e., in the top 25 percent of the acute PM<sub>2.5</sub> vulnerability index) in 2013. White lines correspond to county borders. Gray areas indicate ZIP codes where the majority of the population lives outside of the Commuting Zone or ZIP codes with fewer than 100 beneficiaries.

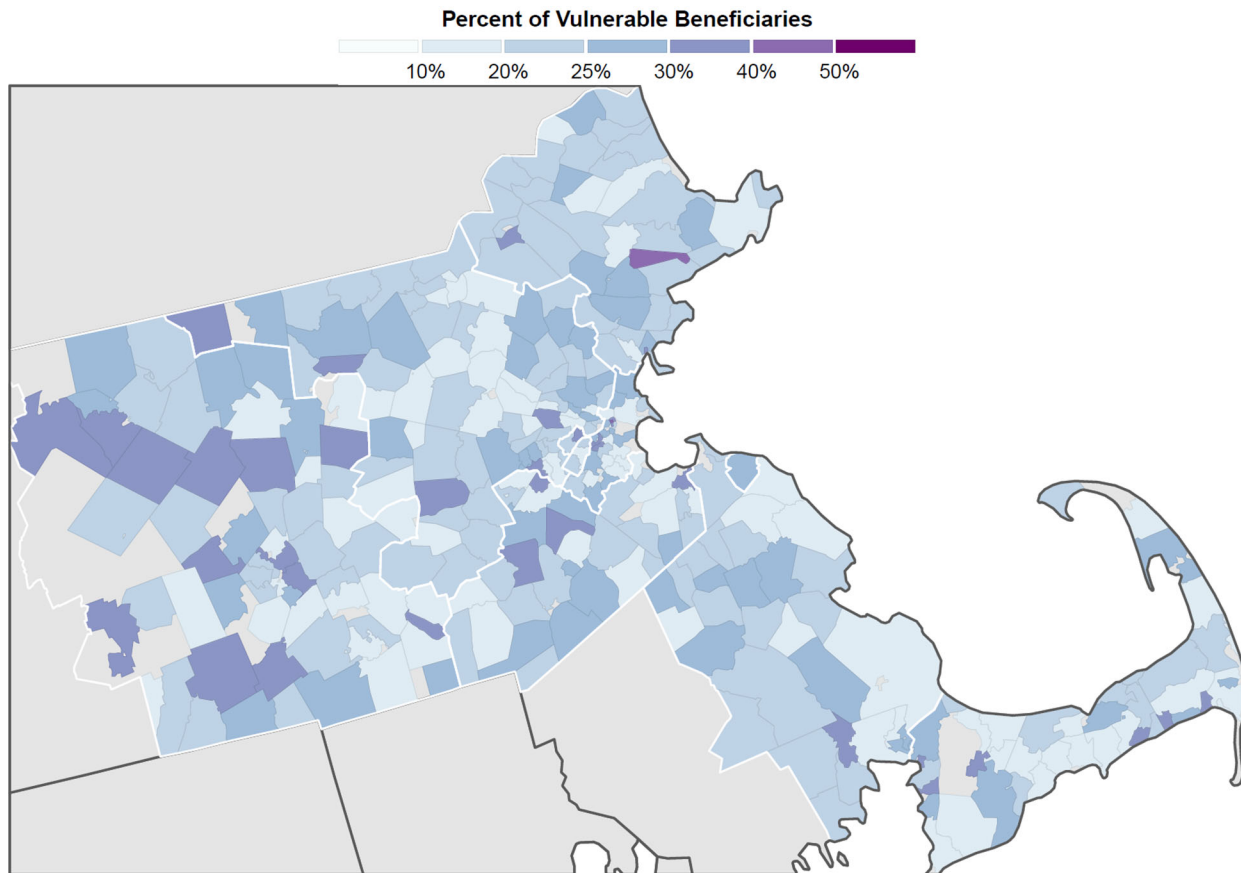


Figure A.4. The map reports ZIP-code-level vulnerability to acute PM<sub>2.5</sub> exposure for all counties in the Commuting Zone containing the Boston-Quincy, MA Metropolitan Division. ZIP code shading indicates the fraction of Medicare beneficiaries in each ZIP code who were vulnerable to acute PM<sub>2.5</sub> exposure (i.e., in the top 25 percent of the acute PM<sub>2.5</sub> vulnerability index) in 2013. White lines correspond to county borders. Gray areas indicate ZIP codes where the majority of the population lives outside of the Commuting Zone or ZIP codes with fewer than 100 beneficiaries.



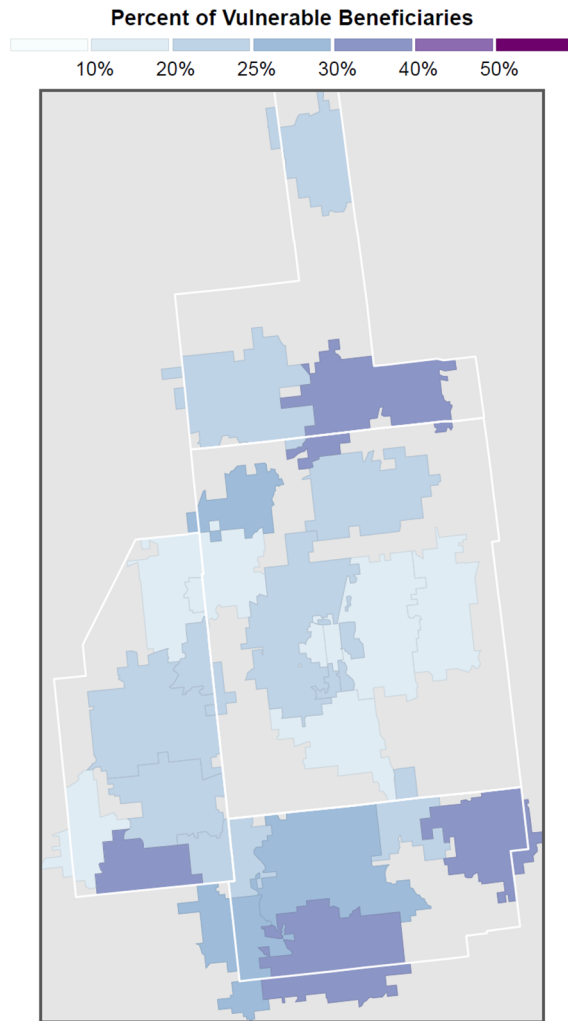


Figure A.5. The map reports ZIP-code-level vulnerability to acute PM<sub>2.5</sub> exposure for all counties in the Commuting Zone containing the Champaign-Urbana, IL Metropolitan Statistical Area. ZIP code shading indicates the fraction of Medicare beneficiaries in each ZIP code who were vulnerable to acute PM<sub>2.5</sub> exposure (i.e., in the top 25 percent of the acute PM<sub>2.5</sub> vulnerability index) in 2013. White lines correspond to county borders. Gray areas indicate ZIP codes where the majority of the population lives outside of the Commuting Zone or ZIP codes with fewer than 100 beneficiaries.

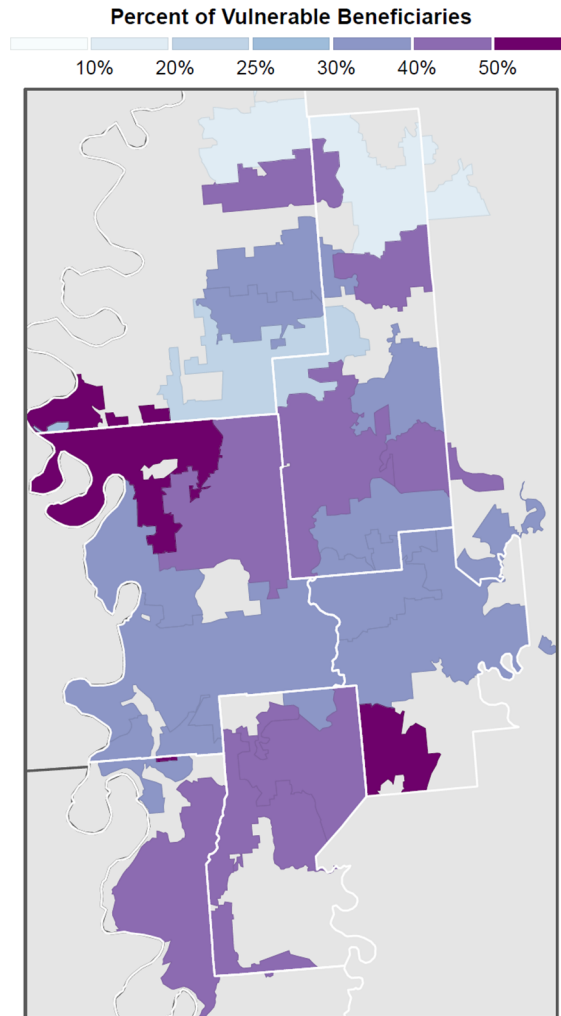


Figure A.6. The map reports ZIP-code-level vulnerability to acute PM<sub>2.5</sub> exposure for all counties in the Commuting Zone containing the Greenville, MS Micropolitan Statistical Area. ZIP code shading indicates the fraction of Medicare beneficiaries in each ZIP code who were vulnerable to acute PM<sub>2.5</sub> exposure (i.e., in the top 25 percent of the acute PM<sub>2.5</sub> vulnerability index) in 2013. White lines correspond to county borders. Gray areas indicate ZIP codes where the majority of the population lives outside of the Commuting Zone or ZIP codes with fewer than 100 beneficiaries.

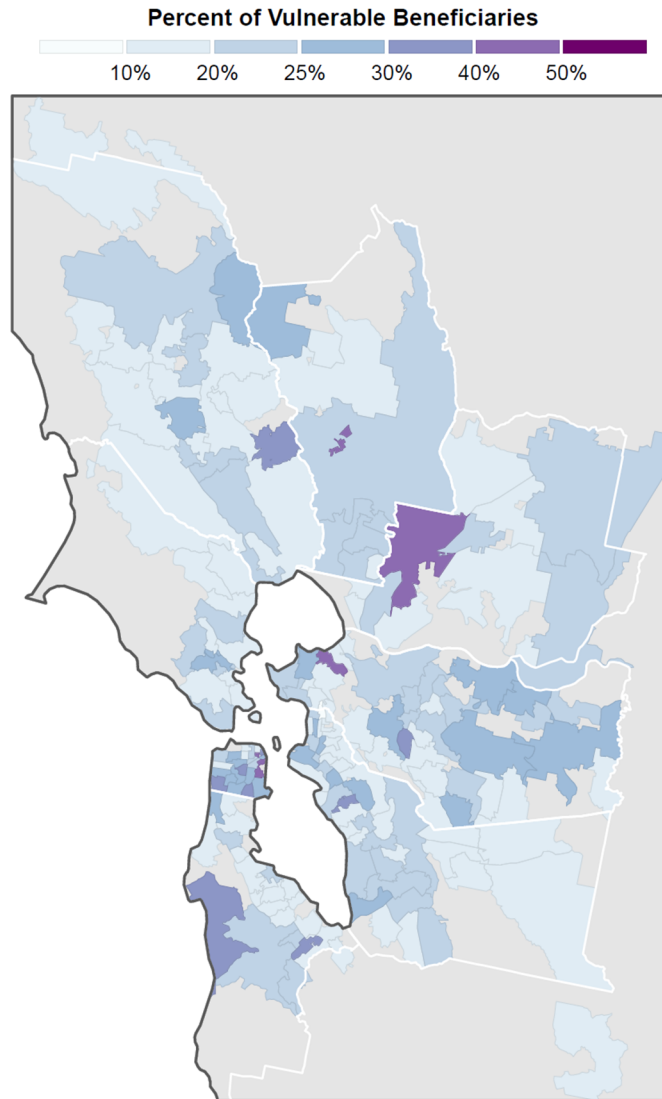


Figure A.7. The map reports ZIP-code-level vulnerability to acute PM<sub>2.5</sub> exposure for all counties in the Commuting Zone containing the Oakland-Fremont-Hayward, CA Metropolitan Division. ZIP code shading indicates the fraction of Medicare beneficiaries in each ZIP code who were vulnerable to acute PM<sub>2.5</sub> exposure (i.e., in the top 25 percent of the acute PM<sub>2.5</sub> vulnerability index) in 2013. White lines correspond to county borders. Gray areas indicate ZIP codes where the majority of the population lives outside of the Commuting Zone or ZIP codes with fewer than 100 beneficiaries.