Supplementary Materials for *Evaluation of the health impacts of the 1990 Clean Air Act Amendments using causal inference and machine learning*

Rachel C. Nethery[1], Fabrizia Mealli[2], Jason D. Sacks[3], Francesca Dominici[1]

[1]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA USA

[2]Department of Statistics, Computer Science, Applications, University of Florence, Florence, Italy

[3]National Center for Environmental Assessment, Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, NC, USA

1 Traditional approaches to health impact assessments for air pollution regulations

Several tools have been developed worldwide that vary in complexity to estimate the potential health implications of changes in air quality including WHO's AirQ+, Aphekom (Improving Knowledge and Communication for Decision Making on Air Pollution and Health in Europe), and U.S. EPA's Environmental Benefits Mapping and Analysis Program – Community Edition (BenMAP – CE) (Goudarzi et al., 2012; Pascal et al., 2013; Sacks et al., 2018). Compared to the other approaches, BenMAP – CE is advantageous because of its flexibility to conduct analyses ranging from local to global in scale (Sacks et al., 2018). As a result of this flexibility, BenMAP-CE is able to combine data from simulated air quality models with pollutant-health exposure-response functions (ERF) to estimate the number of health outcomes prevented due to changes in pollutant exposures within a given year. Air quality modeling softwares, like the ECHAM/MESSy Atmospheric Chemistry-Climate Model (EMAC) (Jöckel et al., 2006) and the Community Multiscale Air Quality Modeling System (CMAQ) (US EPA, 2019), combine emissions inventories, meteorological data, and atmospheric chemistry models to estimate gridded pollution concentrations across large areas. In regulation evaluations, BenMAP-CE uses output from various modeling applications such as those mentioned above to estimate gridded factual (with-regulation) pollution exposures and counterfactual (without-regulation) pollution exposures for a given year.

Pollutant-health ERFs are then used to estimate the number of health events prevented due to an improvement in air quality by using a health effect estimate, i.e. a linear model coefficient capturing the relationship between exposure to an air pollutant and the risk of a health event from a peer-reviewed published epidemiologic study, along with additional inputs including the estimated change in a pollutant concentration (from the air quality modeling software), the baseline incidence rate of the health event, and the size of the population exposed. Our explanation of the ERFs follows the one provided in the supporting materials for the most recent Section 812 Analysis, specifically the document entitled Health

and Welfare Benefits Analyses to Support the Second Section 812 Benefit-Cost Analysis of the Clean Air Act (US EPA, 2011). The units of analysis in the ERF approach are typically grid cells. For each grid cell separately, the ERF estimates Δy, the difference in the health outcome of interest under the factual and counterfactual pollutant levels. The ERF requires the following four inputs: 1) an effect estimate from a previous epidemiologic study relating pollutant exposures to the health outcome (β); 2) a baseline incidence rate for the health event of interest (π_0); the population size in the grid cell (P); and the difference in the counterfactual and factual pollutant estimate for the grid cell (Δx). These inputs are plugged into the following formula:

$$\Delta y = \pi_0 P(e^{\beta \Delta x} - 1)$$

Δy is computed for each grid cell separately, and the results are summarized to obtain the total number of the health event prevented by regulation-attributable changes in the pollutant. This procedure is generally performed for each relevant pollutant/health combination as described below.

In conducting regulation-based health impact analyses, EPA examines pollutant/health outcome combinations for which the evidence base is sufficient to conclude that a causal or likely-to-be causal relationship exists, as discussed in EPA's Integrated Science Assessments (US EPA, 2015), and for which there are published epidemiologic studies available that have examined the exposure-response relationship (to obtain the health effect estimate used by BenMAP-CE). With our proposed approach, we do not rely on this threshold of evidence to dictate the health outcomes evaluated, but instead leverage relationships detected in our observed data for estimation and, if the causal identifying assumptions we lay out in the main manuscript are met, then we are assured that we are capturing causal effects of the pollutant changes on the health outcome. Thus, we are able to analyze effects of the regulation on health outcomes for which the current evidence base is not as rich (e.g., dementia) as is available for some health outcomes (e.g., mortality and cardiovascular effects).

2 Data

2.1 Pollution data

We chose to use factual exposures from the hybrid models, instead of the factual exposure estimates produced for the EPA's Section 812 Analysis using atmospheric chemistry models (ACM), for two primary reasons. First, these hybrid models have been shown to have excellent predictive performance and are validated for investigating pollution-health relationships at the population level (Di et al., 2017a, 2017b). Second, the hybrid model pollution exposures are estimated at much higher spatial resolution than the EPA's factual exposure estimates, which can be advantageous for quantification of exposure-health relationships. Because our analysis will rely on exposure-health relationships in the factual data to estimate counterfactual outcomes (and thereby causal effects), high resolution factual pollution data are essential. However, to give a sense of how the zipcode-level factual pollution exposure from these two sources compare, see Figure 1.

While the two approaches generally estimate similar exposures for both $PM_{2.5}$ and O_3, there are two important takeaways from this comparison. First, there is a substantial amount of noisiness in the relationship between the estimates from the two models. Second, for both pollutants, the ACM estimates more extreme high exposures (in the right tail) compared to the hybrid models. These discrepancies could impact our

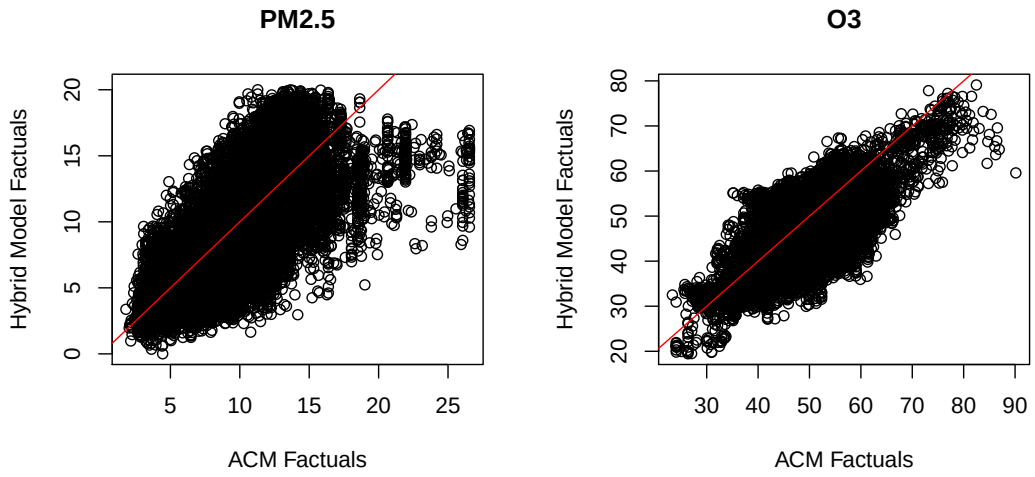

Figure 1: Comparison of factual exposure estimates from hybrid models and those used in the EPA's Section 812 Analysis, which are produced using atmospheric chemistry models (ACM).

analyses in a few different ways. We anticipate that the noisy correspondence in the estimates from the ACM and hybrid models is largely responsible for the fact that, in 30% of zipcodes, our hybrid factual exposure estimates are larger than our ACM counterfactuals. (Note that only 3% of the ACM factual exposures are larger than the ACM counterfactuals). However, in our sensitivity analysis, we have shown that these zipcodes have little impact on our analysis, as we obtain similar TEA estimates when they are excluded. Also, because the ACM exposures may be over-estimated in the right tail, in our setting this could lead to counterfactual exposure estimates that are too extreme in the right tail. Because these extreme counterfactual exposures are unlikely to have matches in the observed data, this could lead to unnecessary excess trimming.

2.2 Medicare data

Here we describe how we constructed counts of each health event from the Medicare data. Mortality counts are created by counting the number of cohort members in each zipcode with dates of death in the range January 1-December 31 of the specified year. In constructing hospitalization counts, we only count each cohort member's first hospitalization (of each type) in the specified year. Cardiovascular hospitalizations are those with a primary diagnosis ICD-9 code 390.xx-459.xx. Dementia hospitalizations are those with either primary or secondary diagnosis Parkinson's Disease (ICD-9: 332.xx), Alzheimer's Disease (ICD-9: 331.0x), or dementia (ICD-9: 290.xx). This definition follows recent work on air pollution and dementia by Kioumourtzoglou et al. (2015).

2.3 Zipcode-level analyses

It is well-known that inference made based on aggregated data, and particularly data aggregated at smaller spatial resolutions, may suffer from the modifiable areal unit problem/ecologic fallacy. Thus, all results

generated by these analyses are inextricably linked to the spatial resolution of the analysis, which in our case is the zipcode level. While analyses at coarser spatial resolutions (e.g., counties), could be less susceptible to the ecologic fallacy, we have chosen to conduct analyses at the zipcode level for several reasons. First, because counties typically cover a much larger spatial area than zipcodes, pollutant exposures within counties can vary substantially and therefore an exposure measure aggregated over the entire county may not adequately capture the exposure of large portions of the county population. Similarly, because the characteristics of populations within counties are often very heterogeneous, we are concerned that aggregating to the county level would compromise our ability to effectively adjust for confounding and capture sources of effect heterogeneity. A final reason for our choice to aggregate to zipcodes is that doing so provides consistency with many recent studies that investigate the impact of pollution exposures on health.

3 Comparison of the TEA with existing causal estimands

There are two features of the TEA that combine to make it unique: a) it accommodates multivariate, continuous treatments; b) it estimates the total number of events attributable to the change in exposure to multiple pollutants, whereas most causal estimands are focused on estimation of average effects. To our knowledge an estimand that captures the number of events attributable to a change in multiple continuous treatments has never been proposed. Recently, a literature has emerged on methods for estimation of population average causal dose-response curves and causal contrasts for a single continuous or multi-valued treatment (Kennedy et al., 2017; Wu et al., 2018). Beyond accommodating only a single treatment variable, the primary limitation of these population average estimands is that they cannot answer our question of interest. We are not concerned about the population average effects of any given level of pollution, but rather the change in outcome that would have occurred in each individual zipcode if its exposures had been the counterfactual no-CAAA exposures rather than the factual exposures. Thus, our estimand must account for the heterogeneity of effects that could be induced by zipcode-specific characteristics, which is ignored in population level estimands. Moreover, these population average estimands require stronger causal identifying assumptions than those needed for our estimand. In particular, the positivity assumption needed for causal dose-response curve estimation requires that each unit have positive probability of receiving any possible exposure level. In finite sample settings this assumption will generally be violated (we will not observe every possible combination of confounding characteristics at every possible exposure level), yet trimming is also not acceptable because it renders the dose-response curve uninterpretable. Without the ability to trim, extrapolation is nearly inevitable. On the other hand, the much weaker positivity assumption for the TEA requires that each combination of counterfactual exposure and confounding characteristics occurs with positive probability in the population. In finite samples, if we do not observe a given combination of exposure and confounding characteristics, we can trim to avoid extrapolation without compromising the utility of our estimate.

In the binary treatment setting, a causal estimand was previously introduced to quantify the number of events attributable to a change in pollution exposure, i.e., above/below some threshold (Baccini et al., 2017). While this estimand is similar in spirit to the TEA, it is of course limited by its focus on a single binary treatment. To our knowledge, all existing causal estimands can accommodate only a single treatment. When estimating separate causal estimands for each pollutant, estimates would be biased unless properly

adjusted for confounding induced by the other pollutant. A variant of our estimand could focus on the events attributable to the changes in each pollutant separately (estimation would need to involve adjustment for confounding by the other pollutants by including them in the Mahalanobis distance). The advantage of taking this approach would be that fewer units would need to be trimmed for each pollutant, since exact matching would need to be performed for only one variable. However, because the estimates for each pollutant would be adjusted for all other pollutant levels, it is not clear how the separate estimates for each pollutant could be combined to quantify the total number of events avoided by the regulation.

4 Asymptotic properties of the matching estimator and bootstrapping

In this section, we discuss the asymptotic properties of our TEA matching estimator, and we provide a bootstrapping approach for obtaining uncertainties for the TEA estimate. We must show that $\hat{\mathbb{E}}(Y|\boldsymbol{X}_i, \boldsymbol{T} \in \Theta_{\boldsymbol{t}_{2i}}) \rightarrow \mathbb{E}(Y|\boldsymbol{X}_i, \boldsymbol{T} \in \Theta_{\boldsymbol{t}_{2i}})$ for all i in order to ensure that $\hat{\tau} \rightarrow \tau$ (recall that asymptotically all units have matches so that we are estimating τ rather than τ^*). Causal inference analyses typically apply matching procedures that identify a common, fixed number of matches (M) for each unit. Abadie and Imbens (2006) showed the consistency of the matching estimator with fixed M and laid out the conditions for $\sqrt{N}$-consistency. However, these results are not applicable to our estimator because we allow for the number of matches to vary across units (M_i) and because our treatment variable is multivariate rather than binary.

To show that each $\hat{\mathbb{E}}(Y|\boldsymbol{X}_i, \boldsymbol{T} \in \Theta_{\boldsymbol{t}_{2i}})$ is consistent, we treat M_i, the number of matches for unit i, as a function of N, so that $M_i \rightarrow \infty$ as $N \rightarrow \infty$. This condition indeed holds as long as the overlap assumption above (A3) is met. In the binary treatment case, Heckman et al. (1998) show that, when M is allowed to go to infinity as N goes to infinity, matching estimators of the average treatment effect on the treated are asymptotically linear and therefore consistent (although not $\sqrt{N}$- consistent). This emerges from the fact that the estimators of the expected counterfactuals can be represented as kernel regression or local polynomial regression, which are asymptotically linear estimators for every set of confounder values, $\boldsymbol{X}_i$. The same result applies to our estimator of $\mathbb{E}(Y|\boldsymbol{X}_i, \boldsymbol{T} \in \Theta_{\boldsymbol{t}_{2i}})$ and its consistency, under the conditions given in Heckman et al. (1998), follows directly from these results. However, the convergence rate is not $\sqrt{N}$.

While some relevant asymptotic normality results (in M) are also provided by Heckman et al. (1998), these may not be reliable in our setting in which M_i is likely to be small for many i. Therefore, we prefer to rely on bootstrapping to obtain inference in this setting. It is a well-known result that the bootstrap fails for matching estimators with fixed M (Abadie and Imbens, 2008), however this failure is precisely due to the fixing of M. The bootstrap is valid for kernel regression (Hall, 1992) and therefore is also valid for our estimator, which can be represented as a kernel regression. However, given the unique structure of our data in this setting, where each unit serves as both (1) a unit that we seek matches for and (2) as a potential match for all other units, how to carry out the bootstrap in practice is not obvious.

We want to obtain the empirical distribution of τ^* using the bootstrap. If we implement the naive bootstrap (resample from the trimmed sample with replacement), some units from the trimmed sample will not appear in the bootstrap sample. Therefore, the set of potential matches changes in the bootstrap sample, meaning that units that had 1 or more matches in the original sample may not have any matches in the

bootstrap sample. Yet, to obtain the empirical distribution of τ^*, we must ensure that no units are trimmed when estimating τ^* in the bootstrap sample. To overcome this issue, we propose the following approach to constructing bootstrap confidence intervals for τ, which has provided reliable results in simulation studies.

1. Resample S units with replacement from the trimmed sample, let $b_1 = 1, ..., B_1$ index these units. These are the units whose estimated causal effects we will sum to produce the bootstrap estimate of τ^*, i.e. the units for which we will seek matches to estimate their counterfactual outcomes.

2. Resample N units with replacement from the entire sample, let $b_2 = 1, ..., B_2$ index these units. This set of units will serve only as potential matches for those from step 1.

3. For each unit $b_1 = 1, ..., B_1$, find matched units $\varphi(b_1)$ in $b_2 = 1, ..., B_2$ using the following approach. Let $\pi(b_1) = \{b_2 \in 1, ..., B_2 : |\boldsymbol{t}_{2b_1} - \boldsymbol{t}_{1b_2}| \prec \boldsymbol{\omega} \ \wedge \ ||\boldsymbol{X}_{b_1} - \boldsymbol{X}_{b_2}|| < \nu\}$. Let $\phi(b_1) = \Big\{j \in 1, ..., B_2 : \sum_{b_2 \neq j} I(||\boldsymbol{t}_{2b_1} - \boldsymbol{t}_{1b_2}|| < ||\boldsymbol{t}_{2b_1} - \boldsymbol{t}_{1j}||) \leq c\Big\}$, i.e. the set of indices of the units b_2 with the c smallest values of $||\boldsymbol{t}_{2b_1} - \boldsymbol{t}_{1b_2}||$, with c a small, pre-specified integer. Define $\rho(b_1) = \text{argmin}_{k \in \phi(b_1)} ||\boldsymbol{X}_{b_1} - \boldsymbol{X}_k||$, then

$$\varphi(b_1) = \begin{cases} \pi(b_1), & \text{if } \pi(b_1) \neq \varnothing \\ \rho(b_1), & \text{otherwise} \end{cases}$$

   With this step, we first seek to find matches for b_1 among the $b_2 = 1, ...B_2$ using the same matching procedure originally applied. In order to ensure that no units are trimmed, if no matches are found with that procedure then we find the unit b_2 with the smallest Mahalanobis distance on confounders among the c units with the smallest Mahalanobis distance on pollutants, and we choose that unit as a single match for b_1.

4. For $b_1 = 1, ..., B_1$, compute $\hat{\mathbb{E}}(Y|\boldsymbol{X}_{b_1}, \boldsymbol{T} \in \Theta_{\boldsymbol{t}_{2b_1}}) = \frac{P_{b_1}}{M_{b_1}} \sum_{k \in \varphi(b_1)} Y_k^*$ (adding a bias correction if one was used to compute the point estimate) and estimate τ^* as above

Repeat this procedure B times to get bootstrap estimates $\{\tau_1^*, ..., \tau_B^*\}$. Then the bootstrap confidence limits are constructed from the percentiles of this empirical distribution, i.e. the 2.5 and 97.5 percentiles are used to create a 95% confidence interval for τ^*.

5 Interpretation of the TEA estimate

The TEA estimate quantifies the evidence of health effects of the regulation that can be ascertained from observed, population-level data, without relying on strong parametric assumptions and model-based extrapolation. What we would ideally like to know is the total number of each health event prevented by the regulation across the entire country; however, we do not have enough support in observed data to inform us about this total. Thus, we are left with two options. The first is to build models using parametric assumptions and use them to extrapolate and estimate totals whose validity are entirely reliant on the subjective parametric modeling assumptions imposed. This is the traditional approach taken by the EPA and other groups. Given the widespread skepticism about the benefits of large-scale environmental regulations, it has become clear that estimates of the health benefits that are heavily model-dependent are vulnerable

to attack and are not, on their own, sufficient to allay concerns about the effectiveness of regulations. More robust evidence is needed. The second option is to extract the maximum amount of information available from observed data about the health effects of the regulation. While we know that the observed data cannot give us a complete picture (due to the lack of support for some of the counterfactual conditions), it provides important insight into the effects that are evident even while making minimal assumptions. This is the approach we take here. We believe the estimated TEA is less vulnerable to attack from skeptics in comparison to the traditional regulation evaluation approach. However, we emphasize that the TEA estimate should not be interpreted as an estimate of overall health effects of the regulation, but instead as an estimate of the health effects that can be robustly estimated using observed data. If we are willing to assume that the regulation had neutral or positive health impacts in all the trimmed units, then the TEA estimate can be interpreted as a sort of minimum number of events avoided by the regulation. The trimming of units with high counterfactual pollution levels is a sacrifice, but it is the price that must be paid to achieve robust, data-driven results, which is our aim here. We advocate that the TEA be considered alongside health benefit estimates using the traditional EPA approach, and that each of their strengths and limitations be appreciated.

6 Details on the generation of pollutants for simulation study

As noted in the main manuscript, for our simulation study we let $N = 4,900$ and $Q = 2$, with one pollutant simulated to mimic $PM_{2.5}$ and the other O_3. The observed pollutants are generated as spatial processes on a 70×70 spatial grid using the `gstat` package in R (Gräler et al., 2016). The vectors of $PM_{2.5}$ and O_3 values (**PM** and **O**) are generated from $\mathbf{PM} \sim MVN(12.18, \mathbf{\Sigma}_1)$ and $\mathbf{O} \sim MVN(0.05, \mathbf{\Sigma}_2)$, where $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_1$ are exponential spatial covariance matrices. The parameters in the exponential covariance matrix are the nugget a, sill s, and range r. In $\mathbf{\Sigma}_1$, they are defined as $a = 3$, $s = 5$, and $r = 10$. In $\mathbf{\Sigma}_2$, they are defined as $a = 0.00005$, $s = 0.00005$, and $r = 10$. O_3 is generated in parts per million and multiplied by 1,000 to correspond to parts per billion.

Note that we have generated $PM_{2.5}$ and O_3 from independent Normal distributions, so that they contain little cross-correlation. In the US, correlation between $PM_{2.5}$ and O_3 is generally low, e.g., in our real data the correlation in the factual $PM_{2.5}$ and O_3 is < 0.3. Given this low degree of cross-correlation in real data, we expect that the performance of the methods on our simulated data will be comparable to their performance in our real data analysis.

7 Robustness of results to exposure error and confounder matching parameter choice

We performed additional sensitivity analyses to evaluate the robustness of our results to potential error in the factual air pollution exposures measures and to the choice of confounder matching parameter ν. As discussed in the main manuscript, we are not aware of any air pollution exposure prediction models that are yet capable of providing reasonable measures of statistical uncertainty. However, to test how sensitive our results may be to factual pollution exposure error, we simulate error of varying degrees, add

it to the factual exposures, and re-fit our models using the error-injected exposures. The structure of the simulated error mimics what we might expect in the real exposure estimates from models built on monitor data, i.e., in densely populated areas with good monitor coverage, errors are small, while in more rural areas where monitors are sparse, errors are larger. To achieve this, we first create an indicator of high population density for each zipcode, discretized at the median empirical population density of 67.65, $W_i = I(popdensity_i > 67.65)$. Then, denoting zipcode aggregated factual $PM_{2.5}$ estimates as PM_i we generate the error-injected exposure as $\tilde{PM}_i = W_i \times \text{TN}(PM_i, (\frac{1}{2}\hat{\sigma}_1\phi)^2) + (1-W_i) \times \text{TN}(PM_i, (\hat{\sigma}_1\phi)^2)$, where $\text{TN}(a, b)$ is a random draw from a truncated normal distribution with mean a, variance b, and truncated below at 0. $\hat{\sigma}_1$ is the empirical standard deviation of the factual $PM_{2.5}$ estimates and ϕ is a parameter used to control the degree of error. Similarly for zipcode aggregated ozone, OZ_i, we generate the error-injected exposure as $\tilde{OZ}_i = W_i \times \text{TN}(OZ_i, (\frac{1}{2}\hat{\sigma}_2\phi)^2) + (1-W_i) \times \text{TN}(OZ_i, (\hat{\sigma}_2\phi)^2)$, where $\hat{\sigma}_2$ is the empirical standard deviation of the factual ozone estimates. We simulate errors under three different specifications of ϕ, emulating scenarios with a small, medium and large degree of error, $\phi = \{0.5,\ 1,\ 2\}$. These reflect the magnitude of error that we might expect to see in real air pollution data. We do not add error to the counterfactual exposures. We conduct analyses of the mortality, dementia hospitalization, and CVD hospitalization outcomes for the year 2000 using these error-injected factual exposures, with all other specifications the same as those described in the primary analysis in the main manuscript. The resulting point estimates and 95% CIs are shown in Figure 2. The results and inference are consistent across each error specification, and these are also consistent with the results of the primary analysis in the main manuscript. This suggests that, even if our factual exposure estimates contain a substantial amount of error, our results are likely to be robust.

We also conduct analyses to test the sensitivity of our results to our choice of ν, the matching parameter that controls the degree of closeness (using the Mahalnobis distance metric) between confounders that is required in order to match two zipcodes. A reasonable selection for ν is crucial for proper confounding adjustment, and ν also plays a role in determining which units are trimmed. In the primary analyses, we specify $\nu = 2.78$, which is approximately the 10^{th} percentile of the Mahalanobis distances between the confounders for each pair of zipcodes in the data. Here we test two alternative values of ν, one smaller and one larger than the value used in the primary analyses. We run analyses for all health outcomes from the year 2000 using the same specifications as in the primary analyses but setting $\nu = 2.30$ and $\nu = 3.32$, corresponding to approximately the 5^{th} and 20^{th} percentile of all Mahalanobis distances, respectively. The results are shown in Figure 3. The results are again consistent with the primary analysis results for both alternative values of ν, suggesting robustness of our results to this choice.

8 EPA's Section 812 Analysis results

Here we provide the estimates of mortalities and CVD hospitalizations prevented from the health impact assessment in the EPA's Section 812 Analysis of the CAAA (US EPA, 2011) (the Section 812 Analysis does not investigate dementia hospitalizations), and we discuss important context that should be considered when comparing these results to our results. While these two analyses seek to estimate the same quantities, the number of each health event prevented due to CAAA-attributable changes in $PM_{2.5}$ and O_3, they take very different statistical approaches that warrant a formal comparison. The strengths and limitations of each approach are detailed in the final two paragraphs of this section. An important piece of context to keep

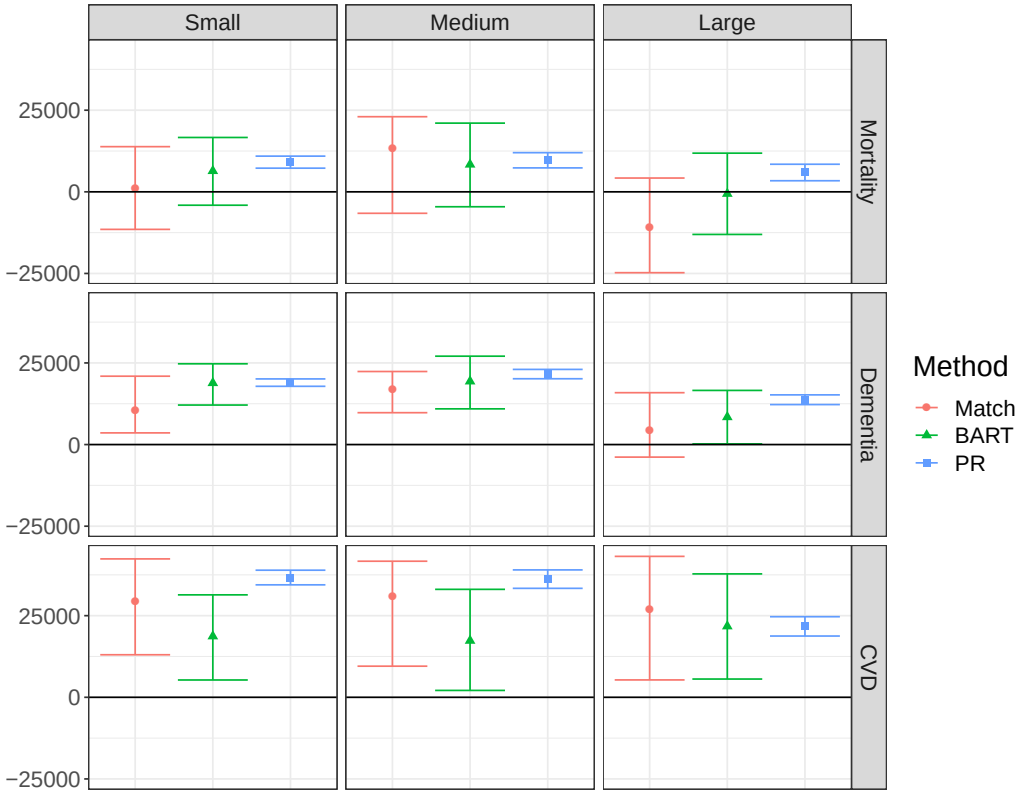

Figure 2: Point estimates and 95% CIs for sensitivity analyses using error-injected factual exposures with a small, medium, and large degree of error.

in mind when directly comparing the results is the different population sizes for which the analyses are conducted, resulting from both the different age groups analyzed and the trimming involved in our method. Due to the different population sizes under study, the estimated number of events from the two methods should not be compared directly but instead compared as a proportion of the underlying population size (or person-time, in the case of the hospitalization analyses). We provide the year-2000 population sizes/person-years, which we refer to as denominators, for the different analyses.

For the year 2000, the Section 812 Analysis estimates that the CAAA-attributable changes in $PM_{2.5}$ prevented 110,000 mortalities in adults age 30+, and the changes in O_3 prevented 1,400 mortalities across all ages (denominator all Americans 30+: 162,603,304). Thus, they estimate that mortality was prevented due to the CAAA in approximately .07% of the population under study. They also estimate that 26,000 CVD hospitalizations were prevented across all ages in 2000 (denominator all Americans: 281,421,906), i.e., CVD hospitalizations were prevented in approximately 0.009% of the population under study due to the CAAA.

In discussing the results of our analyses here, we focus on the BART results for clarity, with BART selected due to its more stable performance across models relative to matching. Our primary analyses (PA) find little evidence of an effect of the CAAA on mortality in the Medicare population in the zipcodes retained after trimming (denominator all Medicare beneficiaries in untrimmed zipcodes: 14,880,606). However,

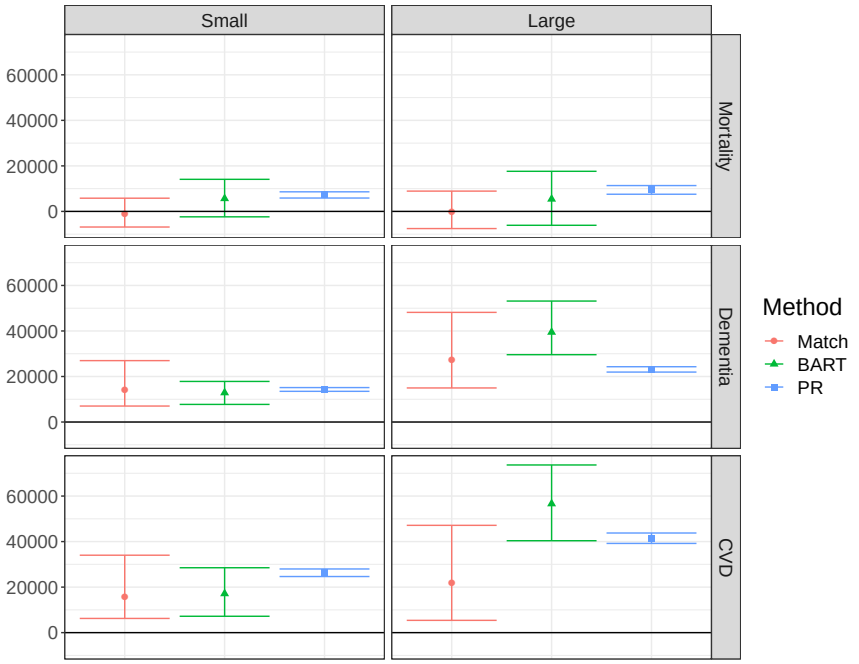

Figure 3: Point estimates and 95% CIs for sensitivity analyses using large and small values of the confounder matching parameter ν ($\nu = 2.30$ and $\nu = 3.32$).

in the sensitivity analysis (SA), BART detects a significant relationship with mortality, estimating 8,033 mortalities prevented by the CAAA in 2000 (denominator all Medicare beneficiaries in untrimmed SA zipcodes: 7,820,770). This corresponds to prevention of mortality in 0.1% of the population under study due to the CAAA. This percentage is consistent with, though slightly larger than, the percentage estimated by the Section 812 Analysis. In the PA, BART estimates that approximately 22,624 CVD hospitalizations were prevented in 2000 (denominator person-years in Medicare FFS in untrimmed zipcodes: 12,766,496), implying that CVD hospitalizations were avoided in 0.2% of the population under study due to the CAAA. In the SA, BART estimates 23,968 avoided CVD hospitalizations (denominator person-years in Medicare FFS in untrimmed SA zipcodes: 6,798,745), corresponding to avoided events in 0.4% of the population under study. Both percentages for CVD hospitalizations are substantially larger than those estimates by the Section 812 Analysis. When comparing the CVD estimates to the Section 812 estimates, note that ERF for CVD in the Section 812 analysis is a pooled estimate from studies that use different sets of ICD-9 codes to define CVD, none of which align perfectly with our set of ICD-9 codes for CVD.

Before considering the strengths and limitations of our method and traditional approach used in the Section 812 Analysis, we make note of the different approaches used by the methods to handle areas whose counterfactual exposures and/or confounders fall outside the range of support of observed data. Our approach removes such areas from the analysis, while the traditional approach uses parametric models to extrapolate the health estimates for these areas. Each approach to handling this issue could be considered a strength or a weakness, depending on one's perspective. Through trimming, our method's results rely on fewer modeling assumptions and are more data-supported, although they reflect a more limited population. The traditional approach is able to produce results for a broader population, but due to the strong modeling assumptions and

extrapolation needed to do so, it is possible that the results could be biased. Considering these trade-offs, both sets of results provide important insights into the health impacts of regulations.

There are several appealing features of the traditional approach used in the Section 812 analysis that are not shared by our approach. The ERFs rely on epidemiologic studies such as the American Cancer Society's Cancer Prevention II study (Pope III et al., 2002) and the Harvard Six Cities Study (Laden et al., 2006), which are built upon individual level data. As discussed in the main manuscript, generally results from individual level data would be preferred to ecologic data, due to increased ability to control for confounding and eliminate noise. Moreover, the Section 812 analysis has the advantage of using factual and counterfactual pollution exposure estimates that were produced in a consistent manner and were designed specifically for that analysis. The spatial coarseness of their exposure estimates is not as problematic in the traditional approach, because they are used as inputs into pre-specified ERFs rather than using them to analyze pollution-health relationships. For reasons discussed in Section 2 of the Supplementary Materials, we believe that the use of the hybrid-model factual pollution estimates is crucial for our analysis; however, our results may suffer from incompatibility in the factual and counterfactual pollution exposure estimates and from the spatial coarseness of the counterfactual estimates.

While the traditional approach has some advantages, it is also overly simplistic in numerous ways that are improved upon by our approach. First, the epidemiologic studies upon which the ERFs are built may not be representative of the population to which inference is made. Our approach improves on this by using real population-level data from the population under study. Second, the ERFs generally impose a linear relationship between the pollutants and health outcomes. Notably, a linear function can lead to extremely high extrapolated estimates of counterfactual outcomes in areas where counterfactual pollution levels are very high (and outside the range of observed pollutant exposures). Third, the long-term pollution exposure and health studies from which the ERFs are obtained rely on a user-specified model form for confounding adjustment, generally within a parametric model, which may lead to insufficient adjustment for confounding because in practice the true forms of these models are typically not known. Fourth, the traditional approach treats $PM_{2.5}$ and O_3 separately and thereby fails to capture potential interactions between them.

9 Additional tables and figures

Table 1: Model forms and parameter values for simulations. Notation generally follows the main manuscript. The k^{th} component of $\boldsymbol{t}_{1i}$ is indexed using the notation $\boldsymbol{t}_{1i_{(k)}}$. $\tilde{X}_{1i}, ..., \tilde{X}_{4i}$ are the four random predictors, with $\tilde{X}_{1i} \sim N(0,1)$, $\tilde{X}_{2i} \sim Exp(1)$, $\tilde{X}_{3i} \sim Unif(0,1)$, and $\tilde{X}_{4i} \sim N(0, 6.25)$.

S-1, Outcome Model	$log(\lambda_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \beta_6 X_{1i} X_{2i} + \beta_7 X_{3i}^2 + \beta_8 exp(X_{4i})(1+exp(X_{4i}))^{-1} + \beta_9 t_{1i_{(1)}} + \beta_{10} t_{1i_{(2)}} + \beta_{11} t_{1i_{(1)}}^2 + \beta_{12} t_{1i_{(1)}} t_{1i_{(2)}} + \beta_{13} t_{1i_{(1)}} t_{1i_{(2)}} \lvert X_{5i} \rvert + \beta_{14} t_{1i_{(1)}} t_{1i_{(2)}} \lvert X_{4i} \rvert + \beta_{15} \tilde{X}_{1i} + \beta_{16} \tilde{X}_{2i} + \beta_{17} \tilde{X}_{3i} + \beta_{18} \tilde{X}_{4i}$ $\beta_0 = -1,\ \beta_1 = 0.0514,\ \beta_2 = -0.0024,\ \beta_3 = 0.002,\ \beta_4 = -0.0007,\ \beta_5 = 0.0074,\ \beta_6 = 0.0436,\ \beta_7 = -0.0001,\ \beta_8 = -0.1472,\ \beta_9 = 0.008,\ \beta_{10} = 0.0001,\ \beta_{11} = 0.0027,\ \beta_{12} = 0.0007,\ \beta_{13} = 0.0002,\ \beta_{14} = 0.0003,\ \beta_{15} = 0.04,\ \beta_{16} = -0.02,\ \beta_{17} = 0.05,\ \beta_{18} = -0.03$
S-2, Outcome Model	$log(\lambda_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \beta_6 X_{1i} X_{2i} + \beta_7 X_{3i}^2 + \beta_8 exp(X_{4i})(1+exp(X_{4i}))^{-1} + \beta_9 t_{1i_{(1)}} + \beta_{10} t_{1i_{(2)}} + \beta_{11} t_{1i_{(1)}}^2 + \beta_{12} t_{1i_{(1)}} t_{1i_{(2)}} + \beta_{13} \tilde{X}_{1i} + \beta_{14} \tilde{X}_{2i} + \beta_{15} \tilde{X}_{3i} + \beta_{16} \tilde{X}_{4i}$ $\beta_0 = 2,\ \beta_1 = 0.0514,\ \beta_2 = -0.0024,\ \beta_3 = 0.002,\ \beta_4 = -0.0007,\ \beta_5 = 0.0074,\ \beta_6 = 0.0136,\ \beta_7 = -0.0131,\ \beta_8 = -0.2472,\ \beta_9 = 0.004,\ \beta_{10} = 0.0001,\ \beta_{11} = 0.007,\ \beta_{12} = 0.0001,\ \beta_{13} = 0.04,\ \beta_{14} = -0.02,\ \beta_{15} = 0.05,\ \beta_{16} = -0.03$
S-3, Outcome Model	$log(\lambda_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \beta_6 t_{1i_{(1)}} + \beta_7 t_{1i_{(2)}} + \beta_8 \tilde{X}_{1i} + \beta_9 \tilde{X}_{2i} + \beta_{10} \tilde{X}_{3i} + \beta_{11} \tilde{X}_{4i}$ $\beta_0 = 5,\ \beta_1 = 0.0014,\ \beta_2 = -0.0243,\ \beta_3 = 0.0496,\ \beta_4 = -0.0372,\ \beta_5 = 0.0238,\ \beta_6 = 0.01,\ \beta_7 = 0.005,\ \beta_8 = 0.04,\ \beta_9 = -0.02,\ \beta_{10} = 0.05,\ \beta_{11} = -0.03$
Confounder Models	$X_{hi} = \boldsymbol{t}_{1i}' \boldsymbol{\alpha}_h + \epsilon_{hi}$ $\boldsymbol{\alpha}_1 = [-0.229\ \ 0.012]',\ \boldsymbol{\alpha}_2 = [-0.221\ \ -0.005]',\ \boldsymbol{\alpha}_3 = [-0.110\ \ -0.004]',\ \boldsymbol{\alpha}_4 = [-0.289\ \ 0.007]',\ \boldsymbol{\alpha}_5 = [0.175\ \ -0.002]',\ \sigma_1^2 = 0.09,\ \sigma_2^2 = 1.04,\ \sigma_3^2 = 9.00,\ \sigma_4^2 = 4.12,\ \sigma_5^2 = 5.42$

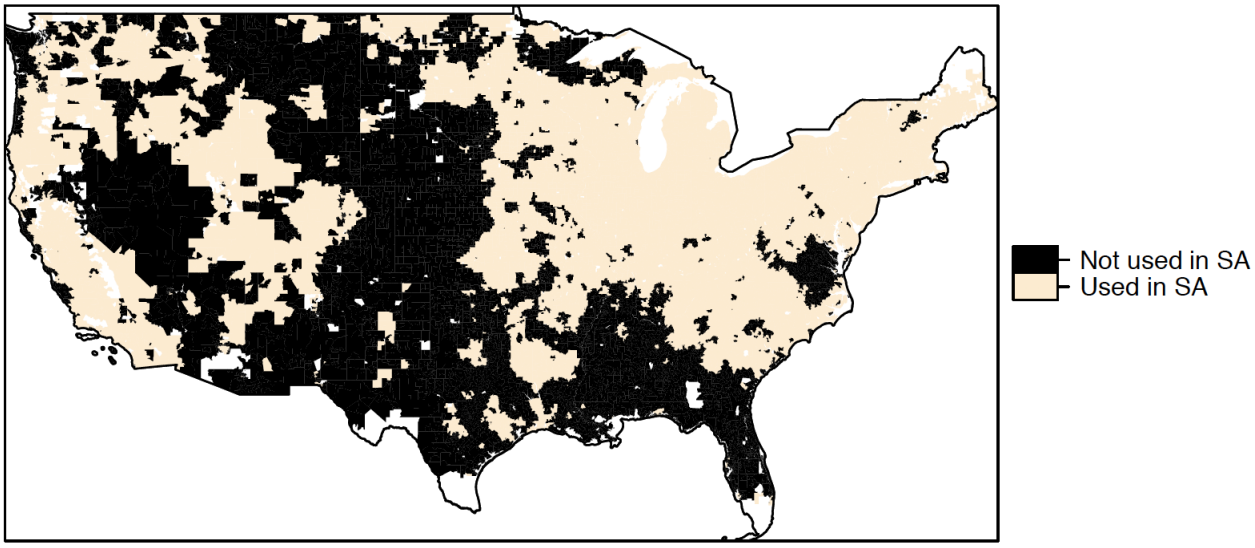

Figure 4: Map of the zipcodes used/not used in the sensitivity analysis. Unused zipcodes are those with one or both factual (with-CAAA) pollution estimates larger than the corresponding counterfactual (no-CAAA).

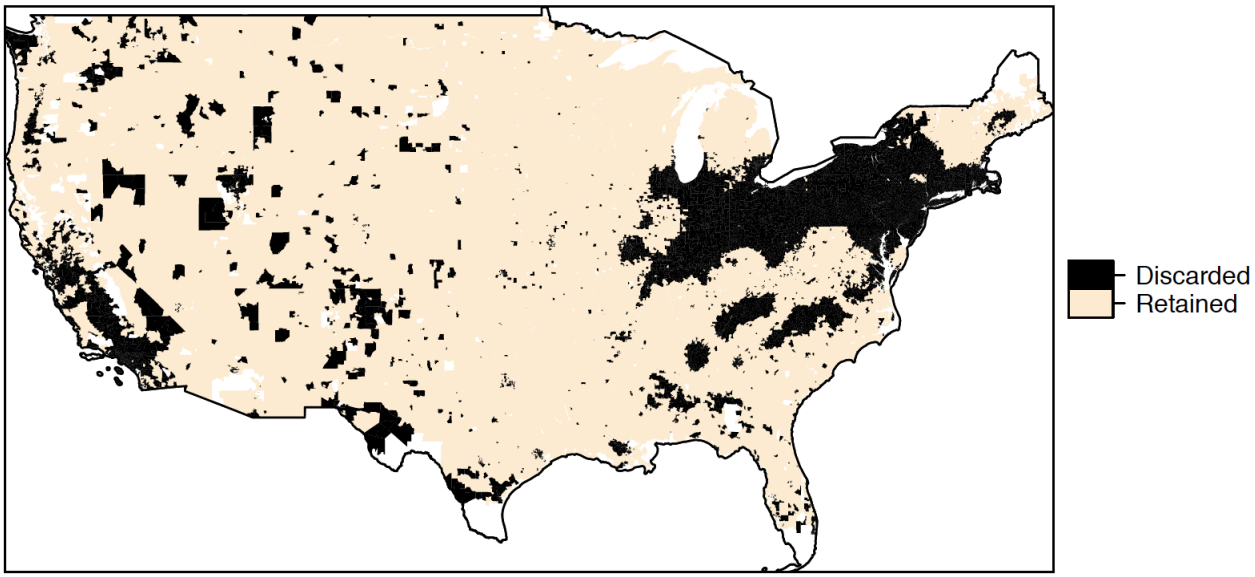

Figure 5: Map of the zipcodes retained and discarded due to trimming in the primary analysis.

Table 2: Average (and standard deviation) of Medicare population size, Medicare health outcome rates, pollutant exposures and confounders in the sensitivity analysis sample, only the discarded/trimmed zipcodes, and the retained/untrimmed zipcodes used for estimation (year 2000 data).

	SA Sample	Discarded Units	Retained Units
Medicare population size	1176.81 (1596.28)	1544.44 (1848.26)	795.52 (1167.31)
FFS person-months	11396.89 (15013.71)	14384.14 (17092.96)	8298.74 (11719.31)
Mortality (rate per 1,000)	52.09 (20.06)	52.63 (20.25)	51.54 (19.85)
Dementia (rate per 1,000)	1.59 (1.26)	1.71 (1.38)	1.46 (1.11)
CVD (rate per 1,000)	6.77 (2.67)	6.83 (2.68)	6.71 (2.65)
Factual $PM_{2.5}$ ($\mu g/m^3$)	10.49 (3.73)	11.44 (3.49)	9.51 (3.71)
Factual O_3 (ppb)	46.76 (6.37)	46.82 (6.74)	46.71 (5.97)
Counterfactual $PM_{2.5}$ ($\mu g/m^3$)	15.51 (4.74)	17.96 (4.25)	12.98 (3.79)
Counterfactual O_3 (ppb)	55.38 (7.27)	58.64 (6.85)	51.99 (6.03)
poverty (proportion)	0.1 (0.09)	0.1 (0.1)	0.11 (0.08)
popdensity (per mi^2)	1440.87 (5169.87)	2491.94 (7014.48)	350.78 (1042.09)
housevalue (USD)	114053.08 (89370.54)	137342.57 (110053.63)	89898.9 (50635.57)
black (proportion)	0.06 (0.14)	0.08 (0.16)	0.05 (0.11)
income (USD)	42338.45 (16999.7)	46345.76 (19848.01)	38182.35 (12092.86)
ownhome (proportion)	0.75 (0.15)	0.72 (0.19)	0.78 (0.1)
hispanic (proportion)	0.05 (0.11)	0.07 (0.14)	0.03 (0.07)
education (proportion)	0.38 (0.18)	0.36 (0.18)	0.39 (0.17)
northeast (proportion)	0.26 (0.44)	0.41 (0.49)	0.09 (0.29)
midwest (proportion)	0.36 (0.48)	0.28 (0.45)	0.45 (0.5)
south (proportion)	0.24 (0.43)	0.18 (0.38)	0.31 (0.46)
west (proportion)	0.14 (0.35)	0.13 (0.34)	0.15 (0.35)

References

Abadie, A. and G. W. Imbens (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica 74*(1), 235–267.

Abadie, A. and G. W. Imbens (2008). On the failure of the bootstrap for matching estimators. *Econometrica 76*(6), 1537–1557.

Baccini, M., A. Mattei, F. Mealli, P. A. Bertazzi, and M. Carugno (2017). Assessing the short term impact of air pollution on mortality: a matching approach. *Environmental Health 16*(7).

Di, Q., L. Dai, Y. Wang, A. Zanobetti, C. Choirat, J. D. Schwartz, and F. Dominici (2017). Association of short-term exposure to air pollution with mortality in older adults. *Jama 318*(24), 2446–2456.

Di, Q., Y. Wang, A. Zanobetti, Y. Wang, P. Koutrakis, C. Choirat, F. Dominici, and J. D. Schwartz (2017). Air pollution and mortality in the Medicare population. *New England Journal of Medicine 376*(26), 2513–2522.

Goudarzi, G., M. Mohammadi, K. Ahmadi Angali, A. Neisi, A. Babaei, B. Mohammadi, Z. Soleimani, and S. Geravandi (2012). Estimation of Health Effects Attributed to NO2 Exposure Using AirQ Model. *Archives of Hygiene Sciences 1*(2), 59–66.

Gräler, B., E. Pebesma, and G. Heuvelink (2016). Spatio-temporal interpolation using gstat. *The R Journal 8*, 204–218.

Hall, P. (1992). On bootstrap confidence intervals in nonparametric regression. *The Annals of Statistics 20*(2), 695–711.

Heckman, J., H. Ichimura, and P. Todd (1998). Matching as an econometric evaluation estimator. *The Review of Economic Studies 65*(2), 261–294.

Jöckel, P., H. Tost, A. Pozzer, C. Brühl, J. Buchholz, L. Ganzeveld, P. Hoor, A. Kerkweg, M. Lawrence, R. Sander, and B. Steil (2006). The atmospheric chemistry general circulation model ECHAM5/MESSy1: consistent simulation of ozone from the surface to the mesosphere. *Atmospheric Chemistry and Physics Discussions 6*(4), 6957–7050.

Kennedy, E. H., Z. Ma, M. D. McHugh, and D. S. Small (2017). Non-parametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 79*(4), 1229–1245.

Kioumourtzoglou, M.-A., J. D. Schwartz, M. G. Weisskopf, S. J. Melly, Y. Wang, F. Dominici, and A. Zanobetti (2015). Long-term $PM_{2.5}$ exposure and neurological hospital admissions in the northeastern United States. *Environmental health perspectives 124*(1), 23–29.

Laden, F., J. Schwartz, F. E. Speizer, and D. W. Dockery (2006). Reduction in fine particulate air pollution and mortality: extended follow-up of the Harvard Six Cities study. *American journal of respiratory and critical care medicine 173*(6), 667–672.

Pascal, M., M. Corso, O. Chanel, C. Declercq, C. Badaloni, G. Cesaroni, S. Henschel, K. Meister, D. Haluza, P. Martin-Olmedo, et al. (2013). Assessing the public health impacts of urban air pollution in 25 European cities: results of the Aphekom project. *Science of the Total Environment 449*, 390–400.

Pope III, C. A., R. T. Burnett, M. J. Thun, E. E. Calle, D. Krewski, K. Ito, and G. D. Thurston (2002). Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *Jama 287*(9), 1132–1141.

Sacks, J. D., J. M. Lloyd, Y. Zhu, J. Anderton, C. J. Jang, B. Hubbell, and N. Fann (2018). The Environmental Benefits Mapping and Analysis Program–Community Edition (BenMAP–CE): A tool to estimate the health and economic benefits of reducing air pollution. *Environmental Modelling & Software 104*, 118–129.

US EPA (2011). Benefits and Costs of the Clean Air Act 1990-2020, the Second Prospective Study. Accessed Online: 2019-05-08.

US EPA (2015). Preamble to the Integrated Science Assessments (ISA). U.S. Environmental Protection Agency, Washington, DC, EPA/600/R-15/067.

US EPA (2019). Community Multiscale Air Quality Modeling System (CMAQ). `doi:10.5281/zenodo.107987`. Accessed: 2019-05-08.

Wu, X., F. Mealli, M.-A. Kioumourtzoglou, F. Dominici, and D. Braun (2018). Matching on generalized propensity scores with continuous exposures. *arXiv preprint arXiv:1812.06575*.