Università di Pavia

DEPARTMENT OF BRAIN AND BEHAVIORAL SCIENCES
Ph.D. in "Psicologia, Neuroscienze e Statistica medica"
Curriculum "Biostatistics"

Estimating the Causal Effect of Low Levels of Fine Particulate Matter on Hospitalization.

Ph.D. Candidate
Riggi Emilia

Supervisor
Prof.re Comelli Mario

A.A. 2014-2017

Summary

ABSTRACT

A growing number of epidemiological studies have provided strong evidence for the adverse health effects of air pollution. However, the association at levels below the U.S. Environmental Protection Agency (EPA) standards (12 μg/m^3 of annual average $PM_{2.5}$) is unclear. In addition, a traditional regression framework does not have a causal interpretation, due to sensitivity to model choice.

Our goal was to recreate an experimental design starting from observational data, in order to strengthen the causal interpretation of the link between low levels of $PM_{2.5}$ and hospital admissions.

One of the major issues in environmental epidemiology is confounding. Using aggregate exposure to $PM_{2.5}$ (two-years prior annual average levels at zip codes level) and all-cause hospitalization rate, we compared the standard regression-based approach, in which the confounders at the zip code level were treated as covariates, to an approach in which zip codes were initially matched according to the confounders and then the effect of $PM_{2.5}$ on health was estimated with regression models restricted to matched zip codes.

We showed that observed confounders widely differed depending on the $PM_{2.5}$ levels. We estimated that, even in very low levels of $PM_{2.5}$, increasing long-term exposure to $PM_{2.5}$ by 1 μg/m^3 causally increased all-cause admissions by 6.2% (95% CI = 3.8%, 8.7%) when the range of $PM_{2.5}$ was 3.50-7.83 μg/m^3 ,9.2% (95% CI = 1.9%, 6.9%) with a range of 7.84-8.65 μg/ m^3 and 12% (95% CI = 4.7%, 19.8%) when the exposure range was 9.37-10.29 μg/ m^3 using nearest-neighbor matching. With Mahalanobis distance matching method we estimated that increasing long-term exposure to PM2.5 by 1 μg/ m^3 causally increased all-cause admissions by 4.7% (95%CI = 2.3%, 7.1%), 10.2% (95%CI = 2.2%, 18.9%) and 16.1% (95%CI = 8.7%, 23.9%) in the same restricted range of $PM_{2.5}$ respectively. In addition, also the analysis with all variables as covariates, showed that

increasing long-term exposure to $PM_{2.5}$ by 1 $\mu g/m^3$, even in very low levels, causally increases all-cause admissions.

Our study was rooted in potential outcomes methods for causal inference that consisted of a design phase that sought using observational data to approximate the design of randomized experiments, where "unexposed" (T = 1) and "exposed" (T = 0) units were balanced with respect to observed confounders; and an outcome analysis phase where the causal effects of adverse health effects to air pollution exposure were estimated. We provided strong evidence of different confounders at each shift in exposure to $PM_{2.5}$ and the developed method was robust to model misspecifications. Last but not least, we showed that long-term exposure to $PM_{2.5}$ were causally associated with all-cause hospitalizations, even for exposure levels not exceeding the U.S. EPA standards, suggesting that adverse health effects occur at low levels of fine particles.

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

- Fine particulate matter (PM2.5)
- US Environmental Protection Agency (EPA)
- National Ambient Air Quality Standards (NAAQS)
- Stable unit treatment value assumption (SUTVA)
- Treatment effect (TE)
- Average treatment effects (ATE)
- Average treatment effect strictly for the treated population (ATT)
- Air Quality System (AQS)
- Socioeconomic status (SES)
- Standardized difference in means (SMD)
- Generalized linear model (GLM)
- Rate ratio (RR)

1. INTRODUCTION

Over the past decade, through epidemiological studies (1–6), long-term exposure to air pollution has been associated with adverse health outcomes. They have identified fine particulate matter ($PM_{2.5}$) as the cause of numerous cases of mortality and morbidity from respiratory to cardiovascular diseases.

Previous studies have generally focused on long-term exposures across the entire range of $PM_{2.5}$ concentrations. In 2012, the US Environmental Protection Agency (EPA) set at 12 $\mu g/m^3$ the National Ambient Air Quality Standards (NAAQS) for the annual average of $PM_{2.5}$. Just few studies (7–12) have showed the health effects of air pollution at levels in accordance with or lower than 12 $\mu g/m^3$.

In addition, the majority of the aforementioned studies were observational and this design does not permit conclusions as whether there is a causal relationship.

The objective of many epidemiological studies is to study the causal effect of exposure to an outcome. As exposure in observational studies is not randomly assigned, confounding is a major threat to the validity of the inference estimates. Confounders are those variables that confound the relationship between exposure and outcome because they are associated with both. . Among the methods available to control for confounding, there are: regression adjustments, stratification on covariates and matching.

The first solution to estimate the association between exposure and outcome of interest is a regression model adjusted for any potential confounders. With

covariance adjustment, a model, usually linear, is fit to the regression of *y* on *x*, and is used to create an adjusted estimate of the exposure effect of the form $\overline{y_1} - \overline{y_2} - \hat{\beta}(\overline{x_1} - \overline{x_2})$ where $\overline{y_1}, \overline{x_1}$ *and* $\overline{y_2}, \overline{x_2}$ are the *y* and *x* means in the exposed and unexposed groups, and $\hat{\beta}$ is the estimate of the slope, β, of *y* on *x*. Furthermore, with this adjustment the balance of the distribution of the confounders across exposure groups is not guaranteed. As a consequence, the exposure groups are not comparable with respect to confounding variables.

In terms of stratification, the effects of confounding can be controlled through stratifying levels of the potential confounder, because groups are produced within which the confounder does not vary. At each level or stratum of the potential confounder, subjects are relatively balanced.

The confounding is controlled when, precisely, the association of interest is estimated within each stratum. Although stratification is a robust method of adjustment for potential confounders, it has problems when there are many potential confounding variables. When the number of confounders rises, each stratum has too few subjects to estimate the association between outcome and exposure reliably.

Formally these techniques can be seen as serving two purposes: to increase the precision of comparisons and to remove initial bias due to *x*. Anyway correlation or association is not the same as causation.

Seeing that the goal of the observational study is to estimate the causal effect of an exposure on an outcome and, the presence of potential confounders due to

not randomized exposure assignment, causal inference methods are the correct approaches.

1.1 POTENTIAL OUTCOME FRAMEWORK

In everyday life, causal language is widely used in an informal way. Investigators are often interested in estimating the effect of exposure on an outcome of interest. Causal inference methods are often employed to address such questions. One core piece of causal inference relies on the potential outcome model. The first time we read about the potential outcome framework was in Neyman paper (13) and then with Fisher (14).

The concept of potential outcome was used exclusively in the context of randomized experiment, not in observational studies. It is only more recently, with Rubin (15), that the framework of potential outcomes was associated with observational study settings. The fundamental notion underlying our approach is that causality is tied to an intervention applied to a unit.

From potential outcome point of view, an individual in a population of interest can be exposed to two alternative states of a cause.

Therefore, a causal statement presumes that, although a unit was exposed to a particular risk, the same unit could have been exposed to an alternative risk.

Give a unit a set of exposure, we associate each exposure-unit pair with a potential outcome. We refer these outcomes as potential outcomes because only one will ultimately be realized and therefore possibly observed: the potential outcome corresponding to the risk actually exposed.

The key assumption of the model is that each individual in the population of interest has a potential outcome under each exposure state, even though each individual can be observed in only one exposure state at any point in time. These what-if potential outcomes are counterfactual in the sense that they exist in theory but are not observed.

There are two important aspects of the definition of causal effect: first, the definition of the causal effect depends on the potential outcomes, but it does not depend on which outcome is actually observed; second, the causal effect is the comparison of potential outcomes, for the same subject, at the same moment in time post-exposure.

1.2 THE CAUSAL EFFECT ESTIMATION

For the estimation and inference of causal effect, we need to compare observed outcomes and because there is only one observed potential outcome per subject, we will need to consider multiple units. More specifically, we must observe multiple units, some exposed and some unexposed. In order to exploit the presence of multiple units, we use the stable unit treatment value assumption (SUTVA) (16). The stability assumption states that exposure applied to one unit do not affect the outcome for another unit and the concept that for each unit there is only a single version of each exposure level.

Just to be in compliance with the literature and be more understandable, as follow we refer to exposure as treatment.

Formalizing this conceptualization, let T be the binary treatment of interest, where T is 0 for “control” or 1 for “treated”, and Y be the continuous outcome. Then, according to Rosenbaum & Rubin (17), each unit, *I*, has two potential outcomes, Y^0 and Y^1. The causal effect of treatment, T, on an outcome, Y, for an observational or experimental unit, *i*, can be defined by comparisons between the outcomes that would have occurred under each of the different treatment possibilities.

In this scenario, the treatment effect (TE) for the unit, *i*, can be defined as the difference between Y_i^1 and Y_i^0 ($TE_i = Y_i^1 - Y_i^0$). Usually, when a participant has been assigned to the treatment condition, the outcome, Y^1, is observed and whereas Y^0 is the unobserved counterfactual outcome, specifically referring to what would have happened to the individual if assigned to the control; on the other hand, for the control unit, Y^0 is observed and Y^1 is counterfactual. Therefore, TE_i is not observed for any unit *i*. This is called the “fundamental problem of causal inference” (18).

It is therefore a problem that at most one of the potential outcomes can be realized and observed.

For the reason that for each unit *i*, only one of Y_i^1 and Y_i^0 are observed because individual units can only receive the treatment or the control but not both, interest lies in estimating the average of individual treatment effects (ATE) defined as:

$$ATE = E(TEi) = E(Y_i^1 - Y_i^0) \qquad \text{(Equation 1)}$$

or the average treatment effect strictly for the treated population, defined as:

$$ATT = E(TE_i \mid T = 1) = E(Y_i^1 - Y_i^0 \mid T_i = 1) \qquad \text{(Equation 2)}$$

The fundamental problem of causal inference is the presence of missing data due to the "assignment mechanism": it determines which units receive with treatments, hence which potential outcomes are observed and, as a consequence, which potential outcomes are missing. This "assignment mechanism" turns out to be the key component in a causal analysis. Fortunately, a classical randomized design, based on each unit having the same probability of receiving each of the possible treatments, can be used to compare treatment and control outcomes.

Randomized experiments have their origins in the work of a statistician Ronald A. Fisher during the 1920s (14). By the definition of Cox and Reid (19), a randomized experiment can be defined as:

"The word experiment is used in a quite precise sense to mean an investigation where the system under study is under the control of the investigator. This means that the individuals or material investigated, the nature of the treatments or manipulations under study and the measurement procedures used are all selected, in their important features at least, by the investigator. By contrast in an observational study some of these features, and in particular the allocation of individuals to treatment groups, are outside the investigator's control."

To be classified as classical randomized experiment, the assignment mechanism requires to be individualistic (the dependence on values of covariates and potential outcomes for other units limited), probabilistic (each experimental unit

has a positive probability of being assigned to the treatment and a positive probability of being assigned to the control), unconfounded (given covariates does not depend on potential outcomes) and controlled by the researcher.

Through this, both the $n_{t=1}$ and $n_{t=0}$ groups are randomly selected and they represent the correspondent treated and control units of the entire population. The same is equivalent for Y_i^1 and Y_i^0. The advantage is that randomization allows balance in potential confounders between these two experimental groups. As a consequence, the control group ($n_{t=0}$), when randomly selected, can behave counterfactually for the treatment group and, vice versa, the randomly selected treated group ($n_{t=1}$) counterfactually for the control group. With this illustration, it is clear why a randomized study is considered the "gold standard" design to determine causal effects of treatment.

The ATE can be easily calculated as the difference between the two marginal means of outcomes ($\overline{Y}_1 - \overline{Y}_0$), respectively, the observed mean when the units are assigned to the treatment condition and the observed mean when the units are assigned to the control condition.

As Neyman demonstrated (13), the difference of the observed means ($\overline{Y}_1 - \overline{Y}_0$) between the $n_{t=1}$ treatment and $n_{t=0}$ control samples is an unbiased estimator of the average treatment effect (ATE).

1.3 APPLICATION TO OBSERVATIONAL STUDIES

The aforementioned definition given by Cox and Reid underlines the main difference between experimental and observational studies.

However, what can we do with observational studies where the units are not randomly selected from the entire population?

As we know from observational studies, exposure is observed, where, on the contrary, with randomized studies, treatment is assigned.

In the causal framework, a comparison of outcomes between exposed and unexposed groups is non-confounded if the two populations are comparable for factors that relate to outcomes. Considering that through an observational study, we do not select units at random from the entire population, a systematic difference between the exposed and unexposed group could be present. These systematic differences take the name of confounders, denoted by a high dimensional vector, W, and, as a result, the outcome, Y, could be affected.

With a randomized experiment through a simple difference between the two marginal means ($\overline{Y}_1 - \overline{Y}_0$), we can estimate the effect of the exposure. Yet, with an observational study where there are indeed confounders, the estimation results tend to be more complicated.

In the regression scenario, the resulting linear regression, when we observe confounders, is:

$$y_i = \alpha + \beta_i T + \delta_i W + \varepsilon_i \qquad \text{(Equation 3)}$$

Adding *W* (*W* generally being a vector of *k* disturbing variables), we improve our model because we better explain the variability of Y, but we cannot say that the coefficient of T is the causal effect of the exposure.

When we talk about estimating causal effect with observational data, it is advisable to mimic a randomized experiment as closely as possible by dividing the sample into exposed and unexposed groups with similar covariate distributions. One may wonder why and the solution is, as mentioned earlier, that a randomized experiment is the gold standard for estimating causal effect, because thanks to the "random assignment" of exposure, the researchers feels confident that all the possible accounted and unaccounted confounders will be equally distributed in the groups.

Therefore, simulating the design of a randomized study, can be achieved by employing matching methods, thanks to well-matched samples, of the original exposed and unexposed groups, reducing bias from covariates.

1.4 MATCHING METHODS

Prior to implementing any methods for estimating causal effects, it is important to conduct a *design phase* of an observational study, during which one can construct a sample such that inferences are more robust and credible. There is one important feature of this initial analysis: it does not involve the outcome data.

The first stage of this phase is to assess the degree of balance in the covariate distribution between exposed and unexposed units. Differences in the

distribution of the measured confounders (W) between these two unit groups would lead to confounding bias.

One of the most common numerical balance diagnostics is the standardized difference in means (SMD), defined as the difference in means of each covariate divided by the standard deviation in the full sample. This provides a scale-free way to assess the differences.

Just to remind, when exposed groups have important covariates that are more than one-quarter or one-half of a standard deviation apart, simple regression methods are unreliable to remove biases associated with differences in covariates.

If the basic samples exhibits a substantial amount of imbalance, we may wish to construct a "subsample" that is characterized by better balance. Such a subsample leads to more robust and thus more credible causal inferences. Therefore, we sequentially match each exposed unit to the closest unexposed units through matching methods.

The main purpose of matching is to replicate a randomized experiment where the only difference in the two groups is exposure.

One assumption of matching methods is "ignorability", that means no unobserved differences between the treatment and control groups, conditional upon the observed covariates (22). Therefore, it is necessary to add in the matching procedure all variables known to be related to both treatment assignment and the outcome. After having determined which covariates to include into the matching, another key concept is to define the "distance", that is

the similarity between two units. We have four definition of distance, D_{ij}, between the units, *i* and *j*, for matching (22):

- Exact, where $D_{ij} = \begin{cases} 0 & if \quad W_i = W_j \\ \infty & if \quad W_i \neq W_j \end{cases}$;
- Mahalanobis, where $D_{ij} = \left(W_i - W_j\right)' \sum^{-1} \left(W_i - W_j\right)$;
- Propensity score, where $D_{ij} = \left|e_i - e_j\right|$, e_k is the propensity score for units, *k*;
- Linear propensity score, $D_{ij} = \left|\log it(e_i) - \log it(e_j)\right|$.

Exact matching is the best approach, though it is not perfectly apt when W is highly dimensional; the same is true for Mahalanobis distance. Exact matching requires that for each unit, *i*, there is the exact matched unit, *j,* and often this leads to many units that are not matched. When there are several covariates, fewer than eight (23,24), and they are normally distributed, Mahalanobis distance is quite appropriate.

In contrast, the propensity score summarizes all covariates into one scalar. It is defined for each unit, *I*, as the probability of being treated given the observed covariates, $e_i\left(W_i\right) = P\left(T_i = 1 \mid W_i\right)$.

One of the easiest and common techniques employed is k:1-nearest-neighbor matching (25), which pairs each treated unit, *i*, and the control unit, *j*, with the smallest distance between them. When there are more treated units than controls, matching can be accomplished with replacement. This means that control units can be utilized as matches or more than one treated unit. This can

often decrease bias because controls that look to be similar to many treated units can be used multiple times. Unfortunately, the estimation of the causal effect becomes more complex because the fact that there is more than one matched control in the matched sample must be accounted for by, for example, employing frequency weights.

The next important step, once having applied matching methods, is to check the quality of the resulting matched samples. Such quality is defined as the balance of the empirical distributions of the full set of covariates (W) in the matched treated and control groups. The absolute standardized differences of means should be less than 0.25 (26).

2. AIM

This study aimed to recreate an experimental design starting from observational data through matching methods, in order to strengthen the causal interpretation of the link between low levels of $PM_{2.5}$ and health outcome using aggregate exposure data (annual average $PM_{2.5}$) and hospital admissions at zip code level.

3. METHODS

3.1 DATA

The long-term (annual 2 years prior the reference date) $PM_{2.5}$ levels and temperatures for the year 2013 were obtained at the monitor level from US Environmental Protection Agency (EPA) Air Quality System (AQS) database, accessible through arepa package (20). In order to have all data at the ZIP code level, the monitors were linked to zip codes.

All-cause hospitalization counts and the total number of people at risk for the year 2013 were obtained at the ZIP code level from billing claims of Medicare enrollees who were fee-for-service Medicare beneficiaries (≥ 65 years of age). For each of the $PM_{2.5}$ monitor locations, we acquired annual numbers of hospitalization admissions and people at risk among the Medicare enrollees residing in each ZIP code with a centroid < 6 miles from a $PM_{2.5}$ monitor location.

We gathered ZIP code–level data on community-level confounding variables, including demographic and socioeconomic (SES) information from the U.S. Census 2010 (21). We averaged values over all ZIP codes with centroids within 6 miles of each $PM_{2.5}$ monitor and assigned the averaged value to each monitor.

Based on that fact that all our sources were at the zip code level, we aggregated them into a unique dataset where at each zip code, we had $PM_{2.5}$ concentration, temperature, U.S Census demographic, SES variables, and hospitalization

outcomes. The list of U.S Census demographic and SES variables are presented as Supplementary material.

3.2 STATISTICAL ANALYSIS

Seen that the aim of our study was to estimate the causal-effect of low levels of air pollution exposure on health outcome, we had two important key phases: the design and outcome analysis.

3.2.1 DESIGN PHASE

During this design phase, the idea was to approximate an observational study into a series of hypothetical randomized experiments. We had two "experimental designs":

1. "ONE SAMPLE" where we divided our sample into two groups based on $PM_{2.5}$ levels. We used a cut-off 9 $\mu g/m^3$. The zip codes with $PM_{2.5}$ <9 $\mu g/m^3$ of exposure were considered "unexposed" (T=0), whereas the others were the "exposed" (T=1).
2. "RESTRICTED SAMPLES" where we divided our sample into ten different bins based on $PM_{2.5}$ concentrations as shown in Figure 1. We used the quantile function because we wanted enough zip codes in each bin. The next step was to recreate a restricted design within zip code characterized by consecutive exposure level. For example, Experiment 1 was the joining of the first bin with the second, Experiment 2 the joining of the third bin with the fourth, and so on. In

each experiment, the zip codes of the bin with lower $PM_{2.5}$ exposure were considered "unexposed" (T=0), whereas the others were the "exposed" (T=1). A depiction of this experimental design is presented in Figure 2.

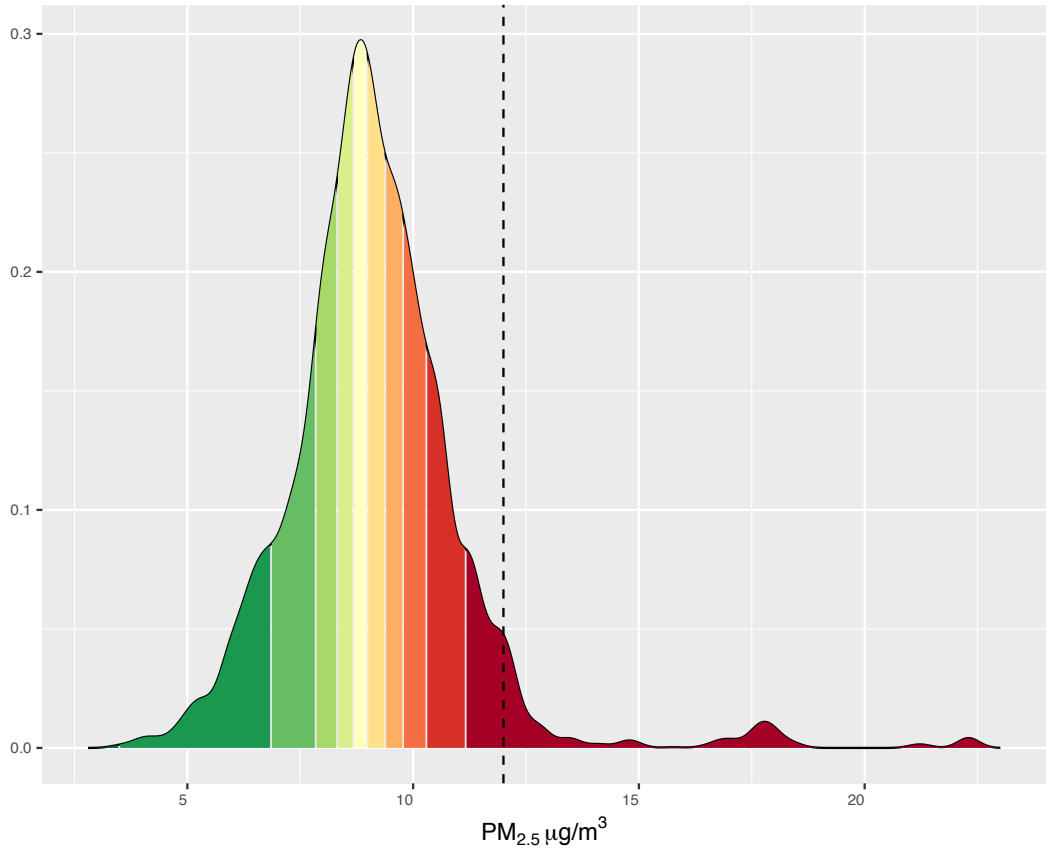


Figure 1. The ten different bins based on exposure to $PM_{2.5}$ levels. The vertical dashed line corresponds to 12 µg/m^3

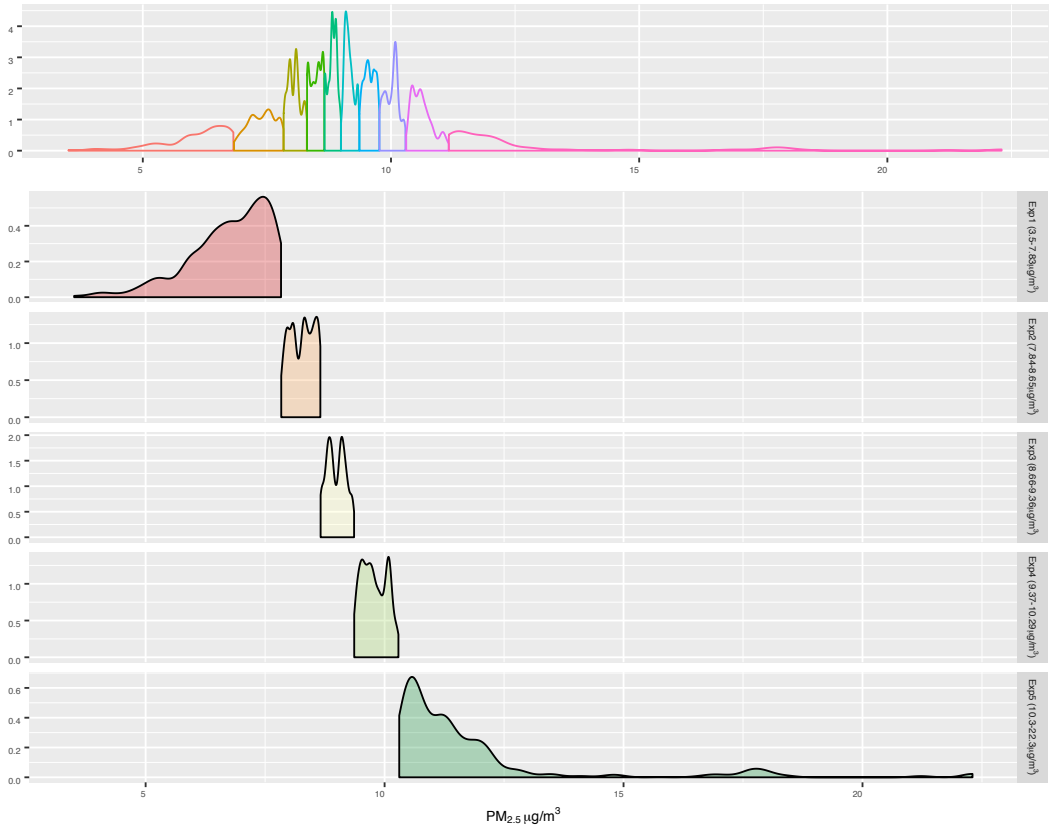


Figure 2. The "RESTRICTED SAMPLES" design: the pairing of bin to design five experiments.

For both experimental designs, we checked for which temperature related variables, demographic variables, and SES, as a full set of variables denoted with W, were unbalanced through the standardized mean difference (SMD) between the exposed and unexposed groups. In addition, we also determined the correlation between these variables (W) and outcomes to discern any potential confounders. In our scenario, a confounder was generally a factor that was simultaneously associated with pollution exposure and health outcomes (specifically, the outcome of interest was all-cause hospital admissions).

To establish if one variable was a potential confounder, we cautiously applied an absolute cutoff of at least 0.1 for both the standardized mean difference and

correlation coefficient. Therefore, those variables turned out to be unbalanced and correlated with the outcomes, and as such were considered confounders.

We employed two different matching methods (17,22) to approximate our observational sample into simulated randomized experiments adjusting for confounders – nearest-neighbor matching and Mahalanobis distance using propensity scoring.

The ultimate goal of the design phase was to construct groups of T = 0 and T = 1 zip codes that were as comparable as possible with respect to W.

After adjusting through matching methods, any model for pollution and health outcomes can be used to estimate causal effects in a manner that is significantly less susceptible to observed confounding (27).

3.2.2 OUTCOME PHASE

The aim of matching methods was to construct a series of experimental design, where groups of T = 0 and T = 1 zip codes were as comparable as possible with respect to W.

After the employment of matching methods, we projected the causal effect of long-term exposure to low levels of $PM_{2.5}$ on a hospitalization outcome, Y.

The outcome variable for our analysis was all-cause hospitalization rate (number of hospitalizations per person-year).

The Poisson distribution is typically used for count data however, we decided to apply a negative binomial regression because through the Poisson regression model, the variance was larger than the mean (called overdispersion) (28,29).

In literature, the negative binomial distribution is presented as a combination of two distributions, giving a combined Poisson-gamma distribution (30). It is important to realize that this distribution is for discrete (integers) and non-negative data.

As hospital admissions were non-negative discrete numbers, and, hence, were not normally distributed, we applied the generalized linear model (GLM) of negative binomial regression (31)

GLM-based negative binomial models are presented in R through the *glm.nb* and *negative.binomial* functions, which are functions in the MASS package (32).

The negative binomial regression model, used to estimate the causal effect on all-cause hospitalization rate of variations for long-term $PM_{2.5}$ exposure, was:

$$Y_i \sim NG(\mu_i, k),\ i = 1,2\ldots,n \qquad \text{(Equation 4)}$$

$$E(Y_i) = \mu_i \text{ and } var(Y_i) = \mu_i + \frac{\mu_i^2}{k} \qquad \text{(Equation 5)}$$

$$\log(\mu_i) = \log(N_i) + \alpha_0 + \beta_1 X_{i1} + \ldots + \beta_q X_{iq} \qquad \text{(Equation 6)}$$

where Y_i and N_i were the number of hospitalizations and the size of the population at risk, as an offset, for the *i*th zip code. In the negative binomial model, μ is the mean and k is the dispersion parameter (31). This model included an offset of the natural logarithm of the number of people at risk in that zip code, taken to be the total number of Medicare enrollees for that zip code in the year 2013. We both fitted models with just $PM_{2.5}$ concentrations as explanatory predictors, and adjusted for all the variables in order to eliminate any residual confounds not accommodated by matching methods and to improve efficiency.

A key point in terms of matching methods is that they are not applied to "compete" with linear regression adjustments. On the contrary, it has been demonstrated that the two methods work best in combination (22). Through this "double robustness," we wanted the exposure effect estimates to be less sensitive to particular outcome model specifications (27). The results were expressed as a rate ratio (RR) estimate. All statistical tests were conducted using an α level of 0.05 and 95% CIs were utilized to measure precision. Statistical analysis was performed with R Statistical software, version 3.3.2 (33). The R code used during the statistical analysis is shown as Supplementary material.

4. RESULTS

In this study, the overall goal was to estimate the causal effect of long-term exposure to low levels of $PM_{2.5}$. For our analysis, data were considered at the zip code level, therefore, for each zip code, we had measures of long-term $PM_{2.5}$, all-cause hospitalization rate and demographic and socioeconomic characteristics. The final dataset contained 5740 zip codes, located as depicted in Figure 3, with an indication of $PM_{2.5}$ concentrations.

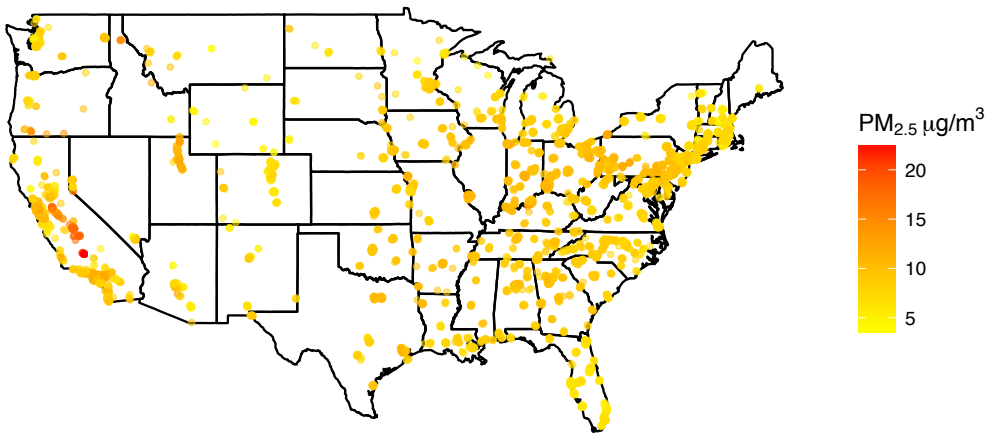


Figure 3. Locations of all 5740 zip codes available for the analysis with average exposure to $PM_{2.5}$ levels in the year 2013.

The median of $PM_{2.5}$ was 8.99 µg/m^3 (range 3.50–22.30 µg/m^3) in the overall sample, in addition 95.6% of our zip codes were designated as "attainment" for $PM_{2.5}$ levels according to the National Ambient Air Quality Standards (NAAQS) (set to 12 µg/m^3).

4.1 ONE SAMPLE

After we divided our sample into two different groups based on $PM_{2.5}$ levels (cut-off 9 µg/m^3), we identified the potential confounders as shown in Figure 4. Section A shows the standardized mean difference (SMD) between the zip codes with $PM_{2.5}$ levels <9 µg/m^3 and those with higher levels. While, section B show the correlation with the outcome.

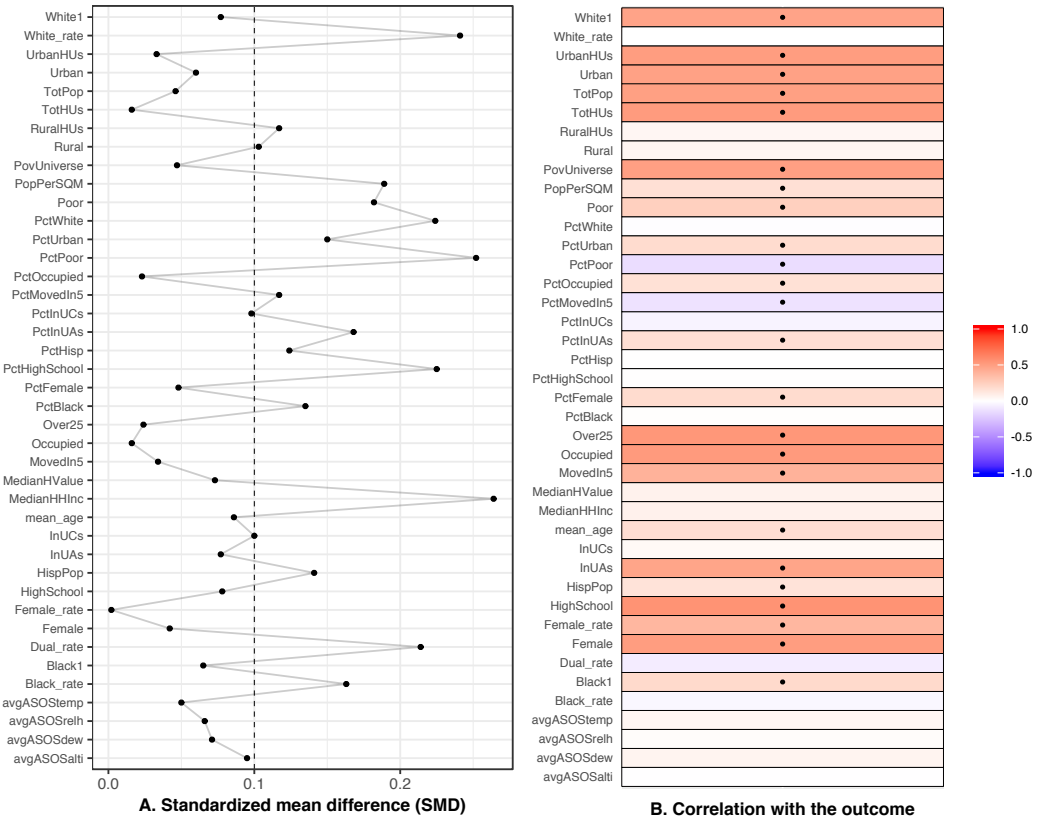


Figure 4. Potential confounders: section A shows the standardized mean difference (SMD) for the full set of variables. Section B visualizes the correlation of our variables with the outcome; in this section those turned out to be correlated with the outcome were highlighted with a dot. For both A and B, we used an absolute cutoff >0.10 to determine variables that were unbalanced and correlated with the outcome.

Those variables turned out to be unbalanced in regard to the exposure and correlated with outcomes, thereby being considered potential confounders. All these variables were used in the next step when we matched the two sets of zip

codes with respect to all the measured potential confounders. We were satisfied when we reached the balance between the “exposed” and “unexposed” units regarding the identified potential confounders as shown in Figure 5.

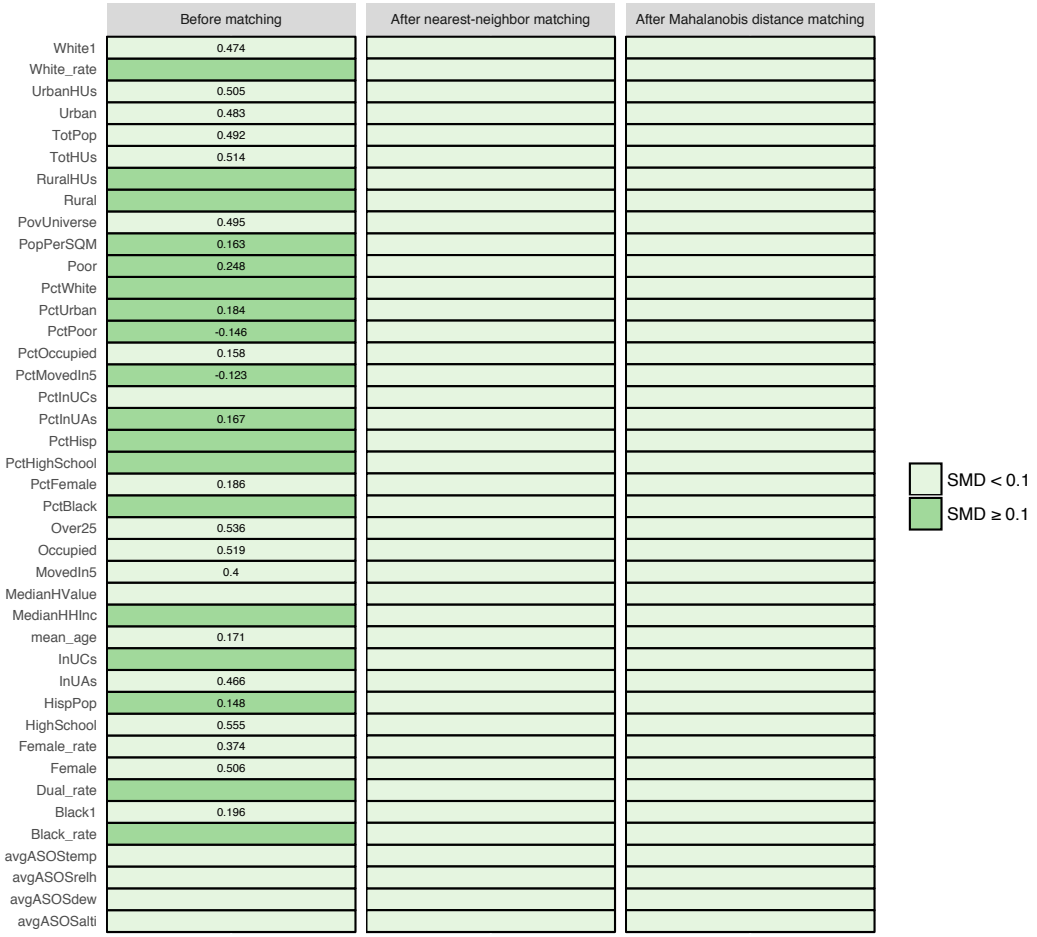


Figure 5. Balance of the identified potential confounders after the application of the employed matching methods. The identified potential confounders were variables that prior to matching were unbalanced (SMD ≥ 0.10) and correlated with outcome (highlighted with correlation coefficient in cells).

With regards to unbalanced sample analysis, we meant the analysis conducted with the sample before application of matching methods. On the contrary, the balanced sample analysis refers to the analysis applied to the sample after matching methods. The results of negative binomial models are shown in table 1.

Table 1. Results of negative binomial regressions with ONE SAMPLE design.

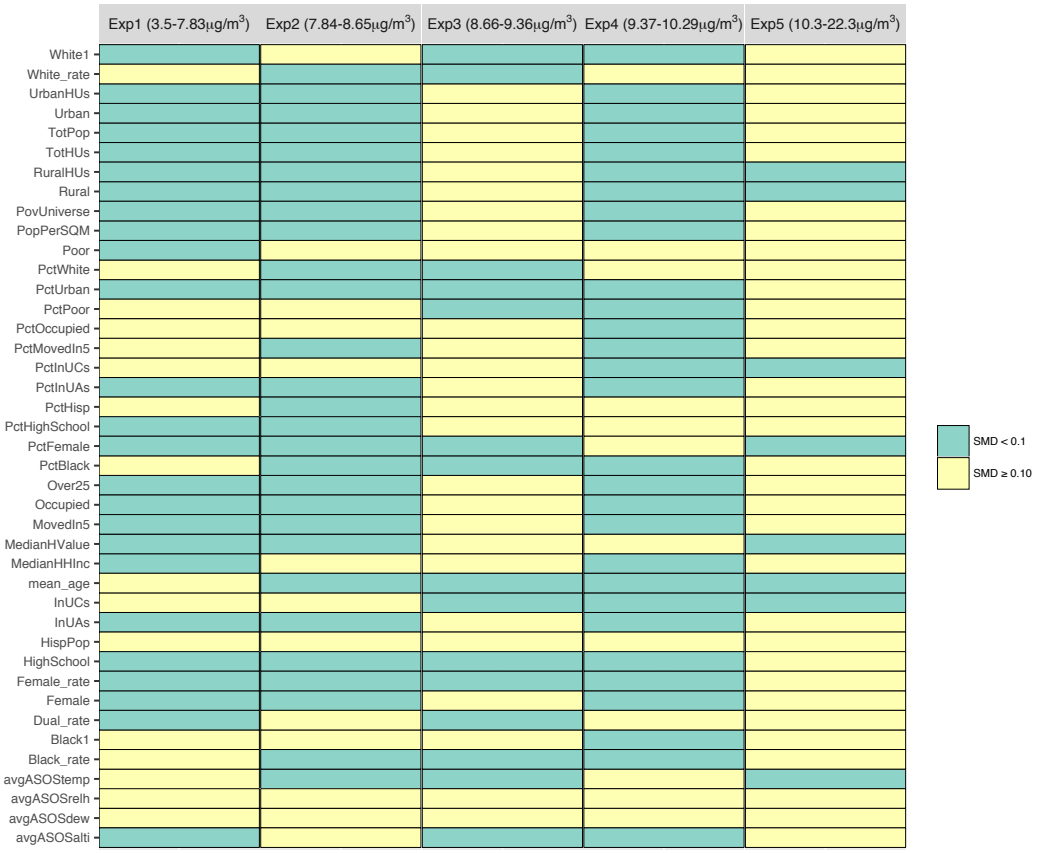
		UNBALANCED SAMPLE ANALYSIS					
		Only $PM_{2.5}$ as explanatory variables			$PM_{2.5}$ + all variables as covariates		
	$PM_{2.5}$ μg/m^3 (range)	RR	95% CI	p-value	RR	95% CI	p-value
One sample (n=5740)	3.50-22.30	1.007	1.003-1.011	0.0002	1.007	1.004-1.010	<0.0001
		BALANCED SAMPLE ANALYSIS					
		Only $PM_{2.5}$ as explanatory variables			$PM_{2.5}$ + all variables as covariates		
	$PM_{2.5}$ μg/m^3 (range)	RR	95% CI	p-value	RR	95% CI	p-value
Nearest-neighbor (n=4478)	3.50-22.30	0.996	0.992-1.001	0.0673	1.006	1.003-1.009	<0.0001
Mahalanobis distance (n=3072)	3.50-22.30	0.995	0.989-0.999	0.0354	1.004	1.000-1.008	0.0472

As mentioned above, we demonstrated that matching methods work better in synergy with linear regressions. Actually, we estimated that increasing long-term exposure to $PM_{2.5}$ by 1 μg/m^3 causally increases all-cause admissions by 0.6% (95% CI = 0.3%, 0.9%) and 0.4% (95% CI = 0%, 0.8%), with Nearest-neighbor and Mahalanobis distance matching methods respectively, when we adjusted for all the variables.

The analysis of balanced sample with only $PM_{2.5}$ as explanatory variable, showed a lower magnitude of the effect, probably due to the residual of some confounds. As consequence, we decided to restrict the levels of $PM_{2.5}$ ("RESTRICTED SAMPLES") in order to eliminate any residual confounds not accommodated with a cut-off 9 μg/m^3 of exposure and to improve efficiency in our estimates.

4.2 RESTRICTED SAMPLES

After we divided our sample into ten different bins based on $PM_{2.5}$ levels, we identified the potential confounders for consecutive deciles of the exposure level distribution (Experiment). As Figure 6 portrays, at different levels of $PM_{2.5}$ exposure we had different sets of potential confounders.



A. Standardized mean difference (SMD)

B. Correlation with the outcome

Figure 6. Section A shows the standardized mean difference (SMD) for the full set of variables, for each experiment, between the zip codes identified as "exposed" and those identified as "unexposed". B visualizes the correlation of our variables with the outcomes. For both A and B, we used an absolute cutoff >0.10 to determine variables that were unbalanced and correlated with the outcome. In B, those that turned out to be correlated with the outcome were highlighted with a dot.

Section A shows the standardized mean difference (SMD) in the unadjusted experiments between the zip codes coded as "exposed" and those identified as "unexposed". It is straightforward to understand that at different exposure levels to $PM_{2.5}$, we had variables that were unbalanced. The same was also valid for the correlation section (B).

Therefore, for each experiment (T = 0, T = 1), those variables turned out to be unbalanced in regard to the exposure and correlated with outcomes, thereby being considered potential confounders. All these variables were used in the

next step when we matched the two sets of zip codes with respect to all the measured potential confounders. We were satisfied when we reached the balance between the “exposed” and “unexposed” units, in each experiment, regarding the identified potential confounders as shown in Figure 7.

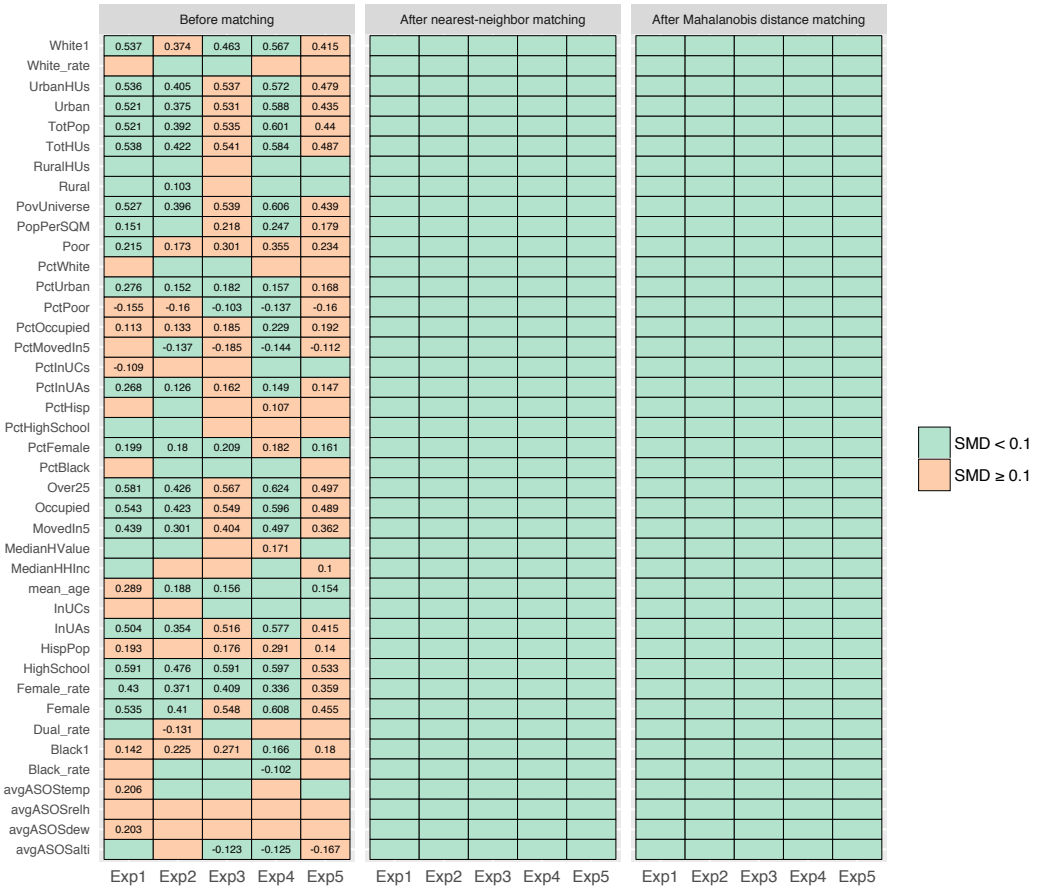


Figure 7. Balance of the identified potential confounders for each experiment after application of the employed matching methods. The identified potential confounders were variables that prior to matching were unbalanced (standardized mean difference (SMD) ≥ 0.1) and correlated with outcomes (were highlighted with correlation coefficient in cells).

From the aforementioned SMD graphing in Figure 7, we concluded that the matching methods did satisfy balancing the covariates between the “exposed” and “unexposed” zip codes. Figure 8 outlines the dimensions for each experiment before and after application of the employed matching methods.

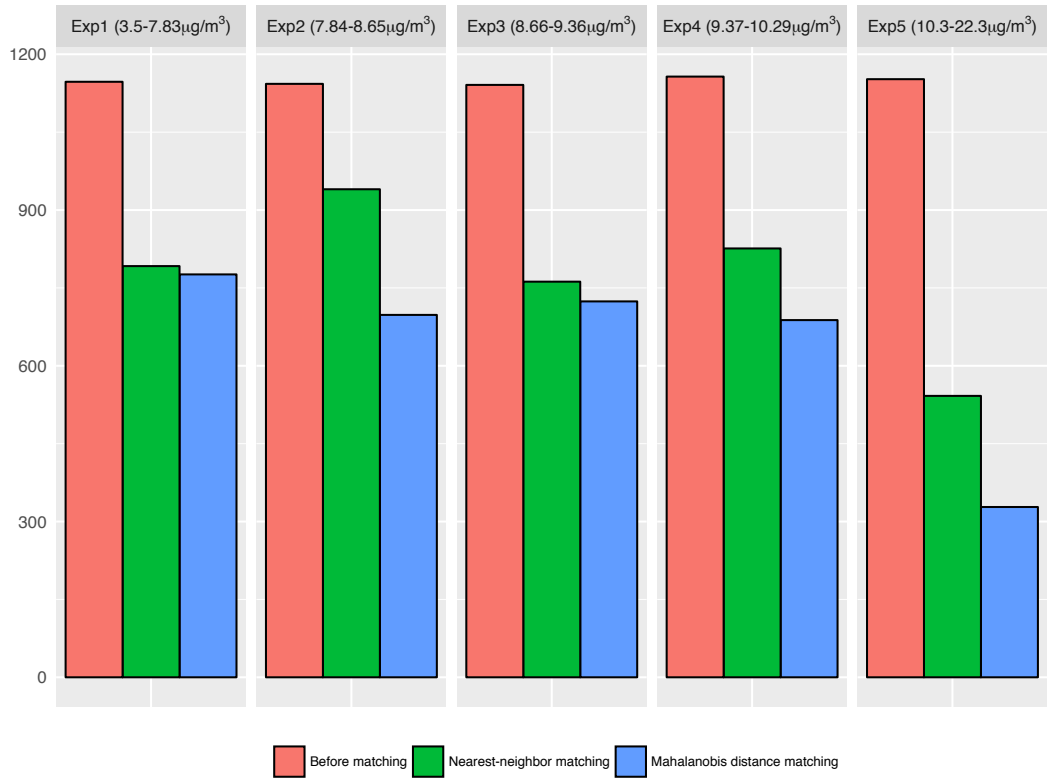


Figure 8. Dimension of the sample for each experiment before and after application of the employed matching methods.

The results of the fitted model are shown in the following table (Table 2). As for "ONE SAMPLE" design, also for the "RESTRICTED SAMPLES" design with regards to unbalanced samples analysis, we meant the analysis conducted with the five experimental samples before application of the employed matching methods, while the balanced samples analysis with the five experimental samples after employing matching methods.

When we applied Nearest-neighbor, we estimated that ,even in very low levels of PM2.5, increasing long-term exposure to $PM_{2.5}$ by 1 μg/m^3 causally increases all-cause admissions by 6.2% (95% CI = 3.8%, 8.7%) , 9.2% (95% CI = 1.9%, 6.9%), 12% (95% CI = 4.7%, 19.8%), in Experiment 1 ($PM_{2.5}$ 3.50-7.83 μg/m^3), 2 ($PM_{2.5}$ 7.84-8.65 μg/m^3), 4 ($PM_{2.5}$ 9.37-10.29 μg/m^3) respectively, while on the contrary

in Experiment 5 ($PM_{2.5}$ 10.30-22.30 μg/m^3) increasing long-term exposure to $PM_{2.5}$ by 1 μg/m^3 causally decreases all-cause admissions by 6.1% (95%CI = -7.7%, -4.4%).

After employing Mahalanobis distance matching method we estimated that increasing long-term exposure to $PM_{2.5}$ by 1 μg/m^3 causally increases all-cause admissions by 4.7% (95%CI = 2.3%, 7.1%), 10.2% (95%CI = 2.2%, 18.9%) and 16.1% (95%CI = 8.7%, 23.9%) in always Experiment 1 ($PM_{2.5}$ 3.50-7.83 μg/m^3), 2 ($PM_{2.5}$ 7.84-8.65 μg/m^3), 4 ($PM_{2.5}$ 9.37-10.29 μg/m^3) respectively, while in Experiment 5 ($PM_{2.5}$ 10.30-22.30 μg/m^3) causally decreases all-cause admissions by 5% (95%CI = -7.3%, -2.6%).

When we move to the analysis with all variables as covariates, we estimated that increasing long-term exposure to $PM_{2.5}$ by 1 μg/m^3 causally increases all-cause admissions by 3% (95%CI = 1.4%, 4.7%), 11.4% (95%CI = 5.6%, 17.5%) and 7.6% (95%CI = 3.1%, 12.2%) in Experiment 1 ($PM_{2.5}$ 3.50-7.83 μg/m^3), 2 ($PM_{2.5}$ 7.84-8.65 μg/m^3), 4 ($PM_{2.5}$ 9.37-10.29 μg/m^3) respectively, while in Experiment 5 ($PM_{2.5}$ 10.30-22.30 μg/m^3) causally decreases all-cause admissions by 2.6% (95%CI = -3.7%, -1.4%), when we used Nearest-neighbor matched samples.

With Mahalanobis distance matched samples, we estimated that increasing long-term exposure to $PM_{2.5}$ by 1 μg/m^3 causally increases all-cause admissions by 2.9% (95%CI = 1.2%, 4.6%), 13% (95%CI = 6.6%, 19.7%) and 11.5% (95%CI = 6.6%, 16.6%) in Experiment 1 ($PM_{2.5}$ 3.50-7.83 μg/m^3), 2 ($PM_{2.5}$ 7.84-8.65 μg/m^3), 4 ($PM_{2.5}$ 9.37-10.29 μg/m^3) respectively, while in Experiment 5 ($PM_{2.5}$ 10.30-22.30 μg/m^3) causally decreases all-cause admissions by 2.2% (95%CI = -3.9%, -0.5%).

Table 2. The RR from the negative binomial regressions for each restricted experiment prior and after applying the matching methods.

Experiment	Exposure $PM_{2.5}$ µg/m^3 (range)	UNBALANCED SAMPLES ANALYSIS					
		Only $PM_{2.5}$ as explanatory variables			$PM_{2.5}$ + all variables as covariates		
		RR	95%CI	p-value	RR	95%CI	p-value
Experiment 1 (n=1147)	3.50 – 7.83	1.061	1.041-1.081	<0.0001	1.033	1.018-1.048	<0.0001
Experiment 2 (n=1143)	7.84 – 8.65	1.073	1.008-1.142	0.027	1.111	1.056-1.168	<0.0001
Experiment 3 (n=1141)	8.66 – 9.36	1.123	1.034-1.220	0.0064	0.964	0.912-1.018	0.2
Experiment 4 (n=1157)	9.37 – 10.29	1.033	0.973-1.10	0.3	1.099	1.056-1.144	<0.0001
Experiment 5 (n=1152)	10.30 – 22.3	0.973	0.965-0.982	<0.0001	0.991	0.984-0.998	0.0083
		BALANCED SAMPLES ANALYSIS					
		Only $PM_{2.5}$ as explanatory variables			$PM_{2.5}$ + all variables as covariates		
	Exposure $PM_{2.5}$ µg/m^3 (range)	RR	95%CI	p-value	RR	95%CI	p-value
Nearest-neighbor matching							
Experiment 1 (n=792)	3.50 – 7.83	1.062	1.038-1.087	<0.0001	1.030	1.014-1.047	0.00031
Experiment 2 (n=940)	7.84 – 8.65	1.092	1.019-1.169	0.013	1.114	1.056-1.175	<0.0001
Experiment 3 (n=762)	8.66 – 9.36	0.950	0.861-1.049	0.3	0.963	0.902-1.029	0.3
Experiment 4 (n=826)	9.37 – 10.29	1.120	1.047-1.198	0.001	1.076	1.031-1.122	0.0007
Experiment 5 (n=542)	10.30 – 22.3	0.939	0.923-0.956	<0.0001	0.974	0.963-0.986	<0.0001
Mahalanobis distance matching							
Experiment 1 (n=776)	3.50 – 7.83	1.047	1.023-1.071	<0.0001	1.029	1.012-1.046	0.0008
Experiment 2 (n=698)	7.84 – 8.65	1.102	1.022-1.189	0.012	1.130	1.066-1.197	<0.0001
Experiment 3 (n=724)	8.66 – 9.36	1.034	0.937-1.141	0.5	1.014	0.949-1.082	0.7
Experiment 4 (n=688)	9.37 – 10.29	1.161	1.087-1.239	<0.0001	1.115	1.066-1.166	<0.0001
Experiment 5 (n=328)	10.30 – 22.3	0.950	0.927-0.974	<0.0001	0.978	0.961-0.995	0.011

To be able to compare the estimates with the ones of "ONE SAMPLE" design, we recombined the entire dataset with the 5 restricted matched samples.

The results of these new models are shown in Table 3. We estimated that increasing long-term exposure to $PM_{2.5}$ by 1 $\mu g/m^3$ causally increases all-cause admissions by 2.3% (95%CI = 1.7%, 2.9%) and by 3.6% (95%CI = 3%, 2.4%), with Nearest-neighbor and Mahalanobis distance matching methods respectively when we added just exposure as explanatory variable.

While, in the analysis with all the variables as covariates, the causal effect of long-term exposure to $PM_{2.5}$ on hospitalization were less but still significant.

Table 3. The results of the negative binomial regressions with the "recombined" dataset.

		Only $PM_{2.5}$ as explanatory variable			**$PM_{2.5}$ + all variables as covariates**		
	$PM_{2.5}$ $\mu g/m^3$ (range)	**RR**	**95% CI**	**p-value**	**RR**	**95% CI**	**p-value**
Nearest-neighbor (n=3862)	3.50-22.30	1.023	1.017-1.029	<0.0001	1.006	1.002-1.010	0.0026
Mahalanobis distance (n=3214)	3.50-22.30	1.036	1.030-1.024	<0.0001	1.014	1.010-1.019	<0.0001

From the "RESTRICTED SAMPLES" design is easy to see that the relationship between exposure $PM_{2.5}$ and all-cause hospital admissions was not linear, probably this was the reason why in the "ONE SAMPLE" analysis, the matched samples were not able to get the causal effect of exposure on all-cause hospital admissions. Therefore, for the recombined samples, we approximated a curve, drawing a natural cubic spline between our findings, to indicate the general, estimated causal effect of increasing long-term exposure to on all-cause hospital admissions, when $PM_{2.5}$ levels is below the National Ambient Air Quality

Standards (NAAQS) (Figure 9). Therefore, we removed zip codes with levels of $PM_{2.5}$ above 12 µg/m^3. For the continuous variable, we defined 4. These 4 knots were 5^{th}, 25^{th}, 75^{th} and 95^{th} percentile of $PM_{2.5}$.

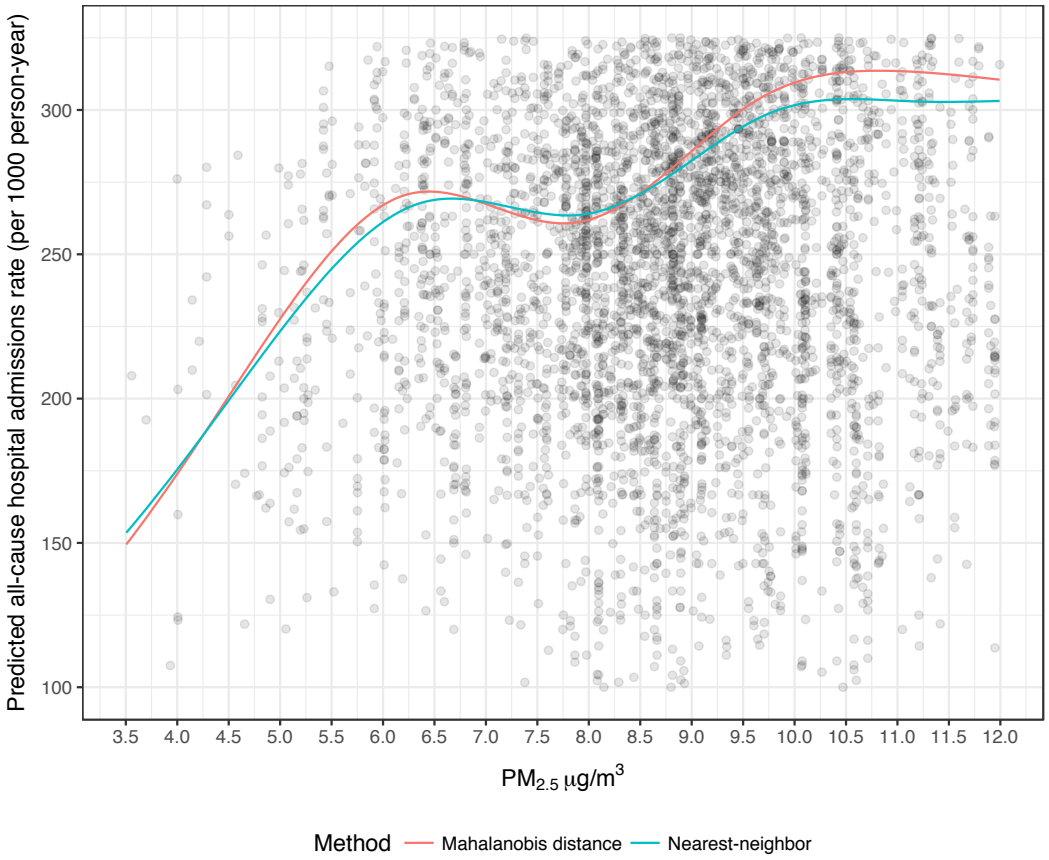


Figure 9. Predicted all-cause hospital admission rate (per 1000 person-year) with a natural cubic spline.

5. DISCUSSION

Our approach was rooted in potential outcomes methods for causal inference that consisted of *a design phase* that sought using observational data to approximate the design of randomized experiments, where "exposed" (T = 1) and "unexposed" (T = 0) units were balanced with respect to observed confounders; and *an outcome analysis phase* where the causal effects of adverse health effects to air pollution exposure were estimated.

We have constructed these hypothetical randomized designs to address the following question: does increasing low level of $PM_{2.5}$ (below the 12 $\mu g/m^3$) causally increase hospitalizations?

To the best of our knowledge, this is the first epidemiological study that provides robust evidence on the adverse health effects of low long-term exposure to $PM_{2.5}$ (2 years average) under the non-confounded assumption by employing matching methods to estimate causal effects.

In our context, confounders were factors that differed between exposed and unexposed groups that also had a relationship with hospitalization outcome. As we reported (Figure 6), observed confounders widely differed depending on $PM_{2.5}$ level. Therefore, at each shift in long-term exposure to $PM_{2.5,}$ we identified different confounders. Our matching method approach was able to group exposed and unexposed zip codes that were most similar on the basis of potential confounders and also estimate the association of being exposed to low levels of $PM_{2.5}$ on all-cause hospital admissions in the causal inference

framework. Comparing the two matching methods, the Nearest-neighbor was a bit better than Mahalanobis distance because it pruned less observations and it reached the balance faster than the other method.

We found that, even in very low levels of PM2.5, increasing long-term exposure to $PM_{2.5}$ by 1 $\mu g/m^3$ causally increased all-cause admissions by 6.2% (95% CI = 3.8%, 8.7%) , 9.2% (95% CI = 1.9%, 6.9%), 12% (95% CI = 4.7%, 19.8%), in Experiment 1 ($PM_{2.5}$ 3.50-7.83 $\mu g/m^3$), 2 ($PM_{2.5}$ 7.84-8.65 $\mu g/m^3$), 4 ($PM_{2.5}$ 9.37-10.29 $\mu g/m^3$) respectively after we applied nearest-neighbor matching.

After employing Mahalanobis distance matching method we estimated that increasing long-term exposure to $PM_{2.5}$ by 1 $\mu g/m^3$ causally increased all-cause admissions by 4.7% (95%CI = 2.3%, 7.1%), 10.2% (95%CI = 2.2%, 18.9%) and 16.1% (95%CI = 8.7%, 23.9%) in always Experiment 1-2-4 respectively.

Also the analysis with all variables as covariates, showed that increasing long-term exposure to $PM_{2.5}$ by 1 $\mu g/m^3$ causally increases all-cause admissions.

Our results were consistent with Makar et. al. study (12), where they found that increasing long-term exposure to $PM_{2.5}$ from levels lower than 12 $\mu g/m^3$ to levels higher than 12 $\mu g/m^3$ increased all-cause admissions. They also found a 15% (95% CI: 8%, 23%) increase in hospitalization rate when $PM_{2.5}$ increased from levels below 8 $\mu g/m^3$ to levels above 8 $\mu g/m^3$.

Generally, several studies continue to demonstrate evidence of an association between long-term exposure to $PM_{2.5}$ and cardiovascular and respiratory morbidity.

Kloog and colleagues (34) used satellite-derived aerosol optical depth (AOD) measurements to predict $PM_{2.5}$ concentrations and reported a 4.22% (95% CI: 1.06,4.75) increase in respiratory hospital admissions associated with a 10 $\mu g/m^3$ rise in long-term $PM_{2.5}$ concentration. Likewise, Neupane et al. (35) observed that long-term exposure to $PM_{2.5}$ was linked with hospitalization for community-acquired pneumonia (OR: 13.64, 95% CI: 1.79,101.01), while Meng et al. (36) noted associations between annual average concentrations of $PM_{2.5}$ and asthma-related emergency department visits and asthma-related hospitalizations. On the contrary, Karr et al. evaluated exposure to $PM_{2.5}$ over an infant's lifetime (0-12 months) and did not observe an association between $PM_{2.5}$ and bronchiolitis hospitalizations.

Other studies (37,38) have reported positive association between short-term exposure to air pollution and cause-specific hospital admissions.

Despite robustness of our method (specifically designed to balance observed confounders), the possibility of residual (such as individual level: "smoking status") and unobserved confounding remained a threat to the validity of our work. Furthermore, having used the Medicare billing claims, where Medicare enrollees are ≥ 65 years of age, was not a convenient sample.

In light of our limitations, our findings should be viewed as preliminary. Here, we have supplied the basis for further exploration in large epidemiologic studies with more detailed information on potential county-level confounders.

In conclusion, we initially demonstrated that at each shift of low long-term exposure to $PM_{2.5}$, there were different confounders. Our findings show that

long-term exposure to $PM_{2.5}$ were causally associated with all-cause hospitalizations, even for exposure levels not exceeding the U.S. EPA standards, suggest that adverse health effects occur at low levels of fine particles.

References

1. Crouse DL, Peters PA, Hystad P, Brook JR, van Donkelaar A, Martin R V, et al. Ambient PM2.5, O_3, and NO_2 Exposures and Associations with Mortality over 16 Years of Follow-Up in the Canadian Census Health and Environment Cohort (CanCHEC). Environ Health Perspect. National Institute of Environmental Health Science; 2015 Nov;123(11):1180–6.
2. Brook RD, Rajagopalan S, Pope CA, Brook JR, Bhatnagar A, Diez-Roux A V., et al. Particulate Matter Air Pollution and Cardiovascular Disease: An Update to the Scientific Statement From the American Heart Association. Circulation. 2010 Jun 1;121(21):2331–78.
3. Dockery DW, Pope CA, Xu X, Spengler JD, Ware JH, Fay ME, et al. An Association between Air Pollution and Mortality in Six U.S. Cities. N Engl J Med. 1993 Dec 9;329(24):1753–9.
4. Samet JM, Zeger SL, Dominici F, Curriero F, Coursac I, Dockery DW, et al. The National Morbidity, Mortality, and Air Pollution Study. Part II: Morbidity and mortality from air pollution in the United States. Res Rep Health Eff Inst. 2000 Jun;94(Pt 2):5-70-9.
5. Dominici F, Peng RD, Bell ML, Pham L, McDermott A, Zeger SL, et al. Fine Particulate Air Pollution and Hospital Admission for Cardiovascular and Respiratory Diseases. JAMA. American Medical Association; 2006 Mar 8;295(10):1127. 7
6. Halonen JI, Lanki T, Yli-Tuomi T, Tiittanen P, Kulmala M, Pekkanen J.

Particulate Air Pollution and Acute Cardiorespiratory Hospital Admissions and Mortality Among the Elderly. Epidemiology. 2009 Jan;20(1):143–53.

7. Crouse DL, Peters PA, van Donkelaar A, Goldberg MS, Villeneuve PJ, Brion O, et al. Risk of Nonaccidental and Cardiovascular Mortality in Relation to Long-term Exposure to Low Concentrations of Fine Particulate Matter: A Canadian National-Level Cohort Study. Environ Health Perspect. 2012 Feb 7;120(5):708–14.

8. Wang Y, Shi L, Lee M, Liu P, Di Q, Zanobetti A, et al. Long-term Exposure to PM2.5 and Mortality Among Older Adults in the Southeastern US. Epidemiology. 2017 Mar;28(2):207–14.

9. Thurston GD, Ahn J, Cromar KR, Shao Y, Reynolds HR, Jerrett M, et al. Ambient Particulate Matter Air Pollution Exposure and Mortality in the NIH-AARP Diet and Health Cohort. Environ Health Perspect. 2016 Sep 15;124:484-90.

10. Pinault L, Tjepkema M, Crouse DL, Weichenthal S, van Donkelaar A, Martin R V., et al. Risk estimates of mortality attributed to low concentrations of ambient fine particulate matter in the Canadian community health survey cohort. Environ Heal. 2016 Dec 11;15(1):18.

11. Shi L, Zanobetti A, Kloog I, Coull BA, Koutrakis P, Melly SJ, et al. Low-Concentration PM2.5 and Mortality: Estimating Acute and Chronic Effects in a Population-Based Study. Environ Health Perspect. 2015 Jun 3;124(1):46–52.

12. Makar M, Antonelli J, Di Q, Cutler D, Schwartz J, Dominici F. Estimating the

Causal Effect of Low Levels of Fine Particulate Matter on Hospitalization. Epidemiology. 2017 Sep;28(5):627–34.

13. Neyman J. On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. Vol. 5, Statistical Science. Institute of Mathematical Statistics; 1990. p. 465–72.
14. Fisher RA. The design of experiments. Oliver and Boyd, Edinburgh; 1935.
15. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol. 1974;66(5):688–701.
16. Rubin DB. Discussion of paper by D. Basu. J Am Stat Assoc. 1980;75:591–3.
17. Rosenbaum PR, B. RD. The central role of the propensity score in observational studies for causal effects. Biometrika. Oxford University Press; 1983;70(1):41–55.
18. Holland PW. Statistics and Causal Inference. J Am Stat Assoc. 1986 Dec;81(396):945–60. Available from:
19. Cox DR (David R, Reid N. The theory of the design of experiments. Chapman & Hall/CRC; 2000. 323 p.
20. Choirat C, Kim C, Zigler C. arepa: An R Package for EPA data retrieving and processing.. Available from: https://github.com/czigler/arepa
21. Bureau of the Census, US Department of Commerce. Census 2010. Washington, DC.
22. Stuart EA. Matching methods for causal inference: A review and a look forward. Stat Sci. NIH Public Access; 2010 Feb 1;25(1):1–21.
23. Rubin DB. Using Multivariate Matched Sampling and Regression

Adjustment to Control Bias in Observational Studies. J Am Stat Assoc. 1979 Jun;74(366):318.

24. Zhao Z. Using Matching to Estimate Treatment Effects: Data Requirements, Matching Metrics, and Monte Carlo Evidence. Rev Econ Stat. MIT Press 238 Main St., Suite 500, Cambridge, MA 02142-1046 USA journals-info@mit.edu ; 2004 Feb;86(1):91–107.

25. Rubin DB. Matching to Remove Bias in Observational Studies. Biometrics. 1973 Mar;29(1):159.

26. Rubin DB. Using Propensity Scores to Help Design Observational Studies: Application to the Tobacco Litigation. Heal Serv Outcomes Res Methodol. Kluwer Academic Publishers; 2001;2(3/4):169–88.

27. Ho DE, Imai K, King G, Stuart EA. Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. Polit Anal. Oxford University Press; 2007 Jan 4;15(3):199–236.

28. Gardner W, Mulvey E, Shaw E. Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models. Psychol Bull. 1995.

29. Zeileis A, Kleiber C, Jackman S, Wien W. Regression Models for Count Data in R. J Stat Softw. 2008;27(8):1–25.

30. Zuur AF, Ieno EN, Walker N, Saveliev AA, Smith GM. Mixed effects models and extensions in ecology with R. New York, NY: Springer New York; 2009. (Statistics for Biology and Health).

31. Hilbe JM. Negative Binomial Regression. Cambridge University Press.

Cambridge University Press; 2007. 576 p.

32. Venables, W. N. & Ripley BD. Modern Applied Statistics with S. Fourth edi. New York, NY: Springer New York; 2002.

33. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2016. Available from: https://www.r-project.org/

34. Kloog I, Coull BA, Zanobetti A, Koutrakis P, Schwartz JD. Acute and Chronic Effects of Particles on Hospital Admissions in New-England. Gravenor MB, editor. PLoS One. Public Library of Science; 2012 Apr 17 ;7(4):e34664.

35. Neupane B, Jerrett M, Burnett RT, Marrie T, Arain A, Loeb M. Long-Term Exposure to Ambient Air Pollution and Risk of Hospitalization with Community-acquired Pneumonia in Older Adults. Am J Respir Crit Care Med. 2010 Jan;181(1):47–53.

36. Meng Y-Y, Rull RP, Wilhelm M, Lombardi C, Balmes J, Ritz B. Outdoor air pollution and uncontrolled asthma in the San Joaquin Valley, California. J Epidemiol Community Heal]. 2010 Feb 1;64(2):142–7.

37. Bell ML, Ebisu K, Peng RD, Walker J, Samet JM, Zeger SL, et al. Seasonal and Regional Short-term Effects of Fine Particles on Hospital Admissions in 202 US Counties, 1999–2005. Am J Epidemiol. US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics, Hyattsville, MD; 2008;168(11):1301–10.

38. Dominici F, Peng RD, Bell ML, Pham L, McDermott A, Zeger SL, et al. Fine Particulate Air Pollution and Hospital Admission for Cardiovascular and Respiratory Diseases. JAMA. 2006 Mar 8;295(10):1127.

Supplementary material

List of U.S Census demographic and SES variables

PctUrban	% Urban population
PctInUAs	% In Urbanized Areas
PctInUCs	% In Urbanized Clusters
PctWhite	% White alone
PctBlack	% Black alone
PctHisp	% Hispanic
PctHighSchool	% High School Grad or GED
MedianHHinc	Median household income
PctPoor	% Persons below poverty
PctFemale	% Female
PctOccupied	% Occupied Housing units
PctMovedIn5	% Moved in last 5 years
MedianHValue	Median Home value
PopPerSQM	Person per Sq mile
TotPop	Total population
Urban	Urban population
InUAs	In Urbanized Areas
InUCs	In Urbanized Clusters
Rural	Rural population
White1	White alone
Black1	Black alone
HispPop	Hispanic
HighSchool	High School Grad or GED
Over25	Over 25 years old
Poor	Poor persons
PovUniverse	Persons for whom poverty status is determined
Female	Female
TotHUs	Total housing units
UrbanHUs	Urban Housing units
RuralHUs	Rural Housing units
Occupied	Occupied housing units
MovedIn5	Moved in last 5 years
avgASOStemp	Average temperature in Fahrenheit
avgASOSdew	Average Dew Point in Fahrenheit
avgASOSalti	Average Pressure altimeter in inches
avgASOSrelh	Average Relative Humidity in %
Mean_age	Average age
Female_rate	Female
White_rate	White
Black_rate	Black
Dual_rate	Dual-eligible beneficiaries

R Code

```
final.imputed<-read.csv(file.choose(), header=T, sep=",",dec=".",
na.strings="NA")
dim(final.imputed)
str(final.imputed$zipcode_R)
names(final.imputed)
library(zipcode)
final.imputed$zipcode_R<-clean.zipcodes(final.imputed$zipcode_R)
str(final.imputed$zipcode_R)


OVERALL<-final.imputed[,-c(16,28,29,50,52:54,59,67:71)]
names(OVERALL)
all.variables<-names(OVERALL)
all.variables
variables<-all.variables[-c(1,16:28,47,56,58)]
length(variables)
variables


#############################################################################
#################################DESIGN PHASE################################
#############################################################################


#########################
####ONE SAMPLE SAMPLE ####
#########################

one.sample<-OVERALL
one.sample$treatment<-ifelse(one.sample$avgPM<9,0,1)
summary(one.sample)
table(one.sample$treatment)


ASD.table<-function(dat, cutoff,index){
  if(is.numeric(index)){
    variables<-names(dat)[index]
  }
  else{
    position<-which(colnames(dat) %in% index)
  }
  variables<-names(dat)[position]
  library(tableone)
  tab<- CreateTableOne(vars = variables, strata = "treatment", data = dat,
test = FALSE)
  OUT<-round(ExtractSmd(tab),3)
  B<-names(OUT[OUT>=cutoff])
  C<-list(OUT,B)
  C
}
tab.RAW<-ASD.table(one.sample,0.10,variables)
tab.RAW
umbalanced.raw<-as.vector(tab.RAW[[2]])
length(umbalanced.raw)

correlated.with.y<-function(dat,var1,var2,cutoff){

  if(is.numeric(var1)){
    var1<-names(dat)[var1]}
  else{var1<-var1}
  if(is.numeric(var2)){
    var2<-names(dat)[var2]}
  else{var2<-var2}
```

```
  corr.mat<-cor(dat[,var1],dat[,var2])
  library(reshape)
  dt<-melt(corr.mat)
  sel<-dt[abs(dt[,3])>cutoff,]
  variables<-sel$X2
  final<-list(corr.mat,sel,variables)
  return(final)
}

correlated.raw<-correlated.with.y(one.sample,"Tot_num.All.cause.admissions",
variables,0.10)
correlated.raw<-as.vector(correlated.raw[[3]])
length(correlated.raw)#24

names(correlated.raw[[1]][1,])

potential.confounder<-function(unbalanced,correlated.y){
  common<-intersect(unbalanced, correlated.y)
  diff1<-setdiff(unbalanced,common)
  diff2<-setdiff(correlated.y, common)
  tomatch<-c(common, diff1,diff2)
  tomatch
}

potential_confounder<-potential.confounder(umbalanced.raw, correlated.raw)
potential_confounder

collinearity<-function(dat,setvar,cutoff){
  corr.mat<-cor(dat[,setvar],dat[,setvar])
  dt<-melt(corr.mat)
  sel<-dt[abs(dt[,3])>cutoff,]
  sel1<-sel[-which(sel[,1]==sel[,2]),]
  nosel<-setdiff(setvar,sel1[,1])
  sel2<-sel1[duplicated(sel1[,3]),]
  sel3<-setdiff(sel2[,1],sel2[,2])
  tomatch<-c(nosel,sel3)
  #
  final<-list("Correlation matrix"=corr.mat,
              "Correlation matrix of selected variables"=sel1,
              "Variables selected"=sel3,"Variables to match"=tomatch)
  return(final)
}

tomatch<-collinearity(one.sample,potential_confounder,0.7)
tomatch<-tomatch$`Variables to match`
length(tomatch)

library(Matching)
library(MatchIt)
nearest_neighbor<-function(dat, index.y,index.x, replace,caliper){
  if(is.numeric(index.x)){
    variables<-colnames(dat)[index.x] }
  else{position<-which(colnames(dat) %in% index.x)}
  variables<-names(dat)[position]
  if(is.numeric(index.y)){
    treatment<-colnames(dat)[index.y]}
  else{position<-which(colnames(dat) %in% index.y)}
  treatment<-names(dat)[position]
  frml<-as.formula(paste(treatment, paste(variables,sep=" ", collapse="
+"),sep="~"))
  ps<-matchit(frml  ,dat,method="nearest",distance="logit",
              reestimate =TRUE, replace=replace, caliper=caliper)
  data.nn<-match.data(ps)
}

mahalanobis.match<-function(dat, index.y, index.x, replace, caliper){
  if(is.numeric(index.x)){
```

```
    variables<-colnames(dat)[index.x]}
  else{position<-which(colnames(dat) %in% index.x)}
  variables<-names(dat)[position]
  if(is.numeric(index.y)){
    treatment<-colnames(dat)[index.y]}
  else{position<-which(colnames(dat) %in% index.y)}
  treatment<-names(dat)[position]
  frml<-as.formula(paste(treatment, paste(variables, sep=" ", collapse="
+"),sep="~"))
  psModel.1 <- glm(frml,family  = binomial(link = "logit"),data    = dat)
  ## Predicted probability of being assigned to TREATED
  dat$pTreated <- predict(psModel.1, type = "response")
  ## Predicted probability of being assigned to CONTROL
  dat$pControl <- 1 - dat$pTreated

  ## Predicted probability of being assigned to the
  ## treatment actually assigned (either TREATED or CONTROL)
  dat$pAssign <- NA
  dat$pAssign[dat[index.y] == "Treated"]    <- dat$pTreated[dat[index.y]   ==
"Treated"]
  dat$pAssign[dat[index.y]  == "Control"]    <-
dat$pControl[dat[index.y]   == "Control"]

  ## Smaller of pTREATED vs pCONTROL for matching weight
  dat$pMin <- pmin(dat$pTreated, dat$pControl)
  listMatch.1 <- Match(Tr       = (dat[index.y] == "1"),      # Need to be in
0,1
                       ## logit of PS,i.e., log(PS/(1-PS)) as matching scale
                       X        = log(dat$pTreated / dat$pControl),
                       ## 1:1 matching
                       M        = 1,
                       ## caliper = 0.2 * SD(logit(PS))
                       caliper  = caliper,
                       replace  = replace,
                       ties     = FALSE,
                       version  = "fast",
                       Weight = 2)
  ## Extract matched data
  dati.Matched.1<-
dat[unlist(listMatch.1[c("index.treated","index.control")]), ]
}


#####MATCHING####

#NEAREST NEIGHBOR
one.sample.near<-
nearest_neighbor(one.sample,"treatment",c(one.sample.tomatch,"avgASOSalti"),F
ALSE,0.02)
tab.NEAR<-ASD.table(one.sample.near,0.10,variables)
tab.NEAR
dim(one.sample.near)
table(one.sample.near$treatment)

#MAHALANOBIS MATCHING
one.sample.maha<-mahalanobis.match(one.sample,"treatment",
c(one.sample.tomatch,"avgASOStemp","avgASOSalti","Dual_rate"),FALSE,0.001)
tab.MAHA<-ASD.table(one.sample.maha,0.10,variables)
tab.MAHA
dim(one.sample.maha)
table(one.sample.maha$treatment)




create.bin<-function(dat,column,upper.cutoff,lower.cutoff){
```

```
  if(is.na(lower.cutoff)){
    bin.1<-subset(dat,dat[,column]<upper.cutoff)
  }
  else {
    if(is.na(upper.cutoff)){
      bin.3<-subset(dat,dat[,column]>=lower.cutoff)
    }
    else{
      bin.2<- subset(dat,dat[,column]<upper.cutoff &
dat[,column]>=lower.cutoff)
    }
  }
  if(is.na(lower.cutoff)){
    return(bin.1)
  }
  else{
    if(is.na(upper.cutoff)){
      return(bin.3)
    }
    else{
      return(bin.2)
    }
  }
}

Q<-quantile(OVERALL$avgPM, probs = seq(0,1,0.10))
Q
bin1<-create.bin(OVERALL,"avgPM",Q[[2]],NA)
dim(bin1)
bin2<-create.bin(OVERALL,"avgPM",Q[[3]],Q[[2]])
dim(bin2)
bin3<-create.bin(OVERALL,"avgPM",Q[[4]],Q[[3]])
dim(bin3)
bin4<-create.bin(OVERALL,"avgPM",Q[[5]],Q[[4]])
dim(bin4)
bin5<-create.bin(OVERALL,"avgPM",Q[[6]],Q[[5]])
dim(bin5)
bin6<-create.bin(OVERALL,"avgPM",Q[[7]],Q[[6]])
dim(bin6)
bin7<-create.bin(OVERALL,"avgPM",Q[[8]],Q[[7]])
dim(bin7)
bin8<-create.bin(OVERALL,"avgPM",Q[[9]],Q[[8]])
dim(bin8)
bin9<-create.bin(OVERALL,"avgPM",Q[[10]],Q[[9]])
dim(bin9)
bin10<-create.bin(OVERALL,"avgPM",NA,Q[[10]])
dim(bin10)

joint_data<-function(subset.1, subset.2) {

  subset.1<-cbind(subset.1, treatment=0)
  subset.2<-cbind(subset.2, treatment=1)
  final<-rbind(subset.1, subset.2)
  return(final)
}


dati.1<-joint_data(bin1,bin2)
length(unique(dati.1$zipcode))


dati.3<-joint_data(bin3,bin4)
length(unique(dati.3$zipcode))


dati.5<-joint_data(bin5,bin6)
```

```
length(unique(dati.5$zipcode))

dati.7<-joint_data(bin7,bin8)
length(unique(dati.7$zipcode))

dati.9<-joint_data(bin9,bin10)
length(unique(dati.9$zipcode))



dati.1$Experiment<-rep("Exp1",times=dim(dati.1)[1])
dati.3$Experiment<-rep("Exp3",times=dim(dati.3)[1])
dati.5$Experiment<-rep("Exp5",times=dim(dati.5)[1])
dati.7$Experiment<-rep("Exp7",times=dim(dati.7)[1])
dati.9$Experiment<-rep("Exp9",times=dim(dati.9)[1])

dati.1$bin.name<-ifelse(dati.1$treatment==0,1,2)
dati.3$bin.name<-ifelse(dati.3$treatment==0,3,4)
dati.5$bin.name<-ifelse(dati.5$treatment==0,5,6)
dati.7$bin.name<-ifelse(dati.7$treatment==0,7,8)
dati.9$bin.name<-ifelse(dati.9$treatment==0,9,10)




##############################################
########## EXPERIMENTS SAMPLE ###############
##############################################
#Umbalanced variables
tab.RAW.exp<-lapply(list(dati.1,dati.3,dati.5,dati.7,dati.9),function(x)
ASD.table(x, 0.10, variables))

umbalanced.raw.1<-tab.RAW.exp[[1]][[2]]
umbalanced.raw.3<-tab.RAW.exp[[3]][[2]]
umbalanced.raw.5<-tab.RAW.exp[[5]][[2]]
umbalanced.raw.7<-tab.RAW.exp[[7]][[2]]
umbalanced.raw.9<-tab.RAW.exp[[9]][[2]]


#Correlated with y
correlated.raw.exp<-
lapply(list(dati.1,dati.3,dati.5,dati.7,dati.9),function(x)
correlated.with.y(x,"Tot_num.All.cause.admissions",variables,0.10))

correlated.raw.1<-
as.vector(correlated.raw.exp[[1]][[3]]);length(correlated.raw.1)
correlated.raw.3<-
as.vector(correlated.raw.exp[[3]][[3]]);length(correlated.raw.3)
correlated.raw.5<-
as.vector(correlated.raw.exp[[5]][[3]]);length(correlated.raw.5)
correlated.raw.7<-
as.vector(correlated.raw.exp[[7]][[3]]);length(correlated.raw.7)
correlated.raw.9<-
as.vector(correlated.raw.exp[[9]][[3]]);length(correlated.raw.9)



#Selection of variables unbalanced and correlated with y
potential.1<-potential.confounder(umbalanced.raw.1, correlated.raw.1)
potential.3<-potential.confounder(umbalanced.raw.3, correlated.raw.3)
potential.5<-potential.confounder(umbalanced.raw.5, correlated.raw.5)
potential.7<-potential.confounder(umbalanced.raw.7, correlated.raw.7)
potential.9<-potential.confounder(umbalanced.raw.9, correlated.raw.9)
```

```
#Selection of variables to match no collinearity
tomatch.1<-collinearity(dati.1,potential.1,0.7)
tomatch.1<-tomatch.1$`Variables to match`
tomatch.3<-collinearity(dati.3,potential.3,0.7)
tomatch.3<-tomatch.3$`Variables to match`
tomatch.5<-collinearity(dati.5,potential.5,0.7)
tomatch.5<-tomatch.5$`Variables to match`
tomatch.7<-collinearity(dati.7,potential.7,0.7)
tomatch.7<-tomatch.7$`Variables to match`
tomatch.9<-collinearity(dati.9,potential.9,0.7)
tomatch.9<-tomatch.9$`Variables to match`

Invert.treatment<-function(dat){
  treatment2<-ifelse(dat$treatment==1,0,1)
}
library(MatchIt)
library(Matching)

#EXP1
table(dati.1$treatment)
dati.1$treatment<-Invert.treatment(dati.1)
exp.1.nearest<-
nearest_neighbor(dati.1,"treatment",c(tomatch.1,"PctWhite","Black_rate"),FALS
E, 0.2)
tabNEAREST.1<-ASD.table(exp.1.nearest,0.10,variables)
tabNEAREST.1
dim(exp.1.nearest)
table(exp.1.nearest$treatment)

exp.1.maha<-
mahalanobis.match(dati.1,"treatment",c(tomatch.1,"PctWhite","Black_rate"),FAL
SE, 0.8)
tabMAHALANOBIS.1<-ASD.table(exp.1.maha, 0.10,variables)
tabMAHALANOBIS.1
dim(exp.1.maha)
table(exp.1.maha$treatment)


#EXP3
exp.3.nearest<-nearest_neighbor(dati.3,"treatment",tomatch.3,FALSE, 0.008)
tabNEAREST.3<-ASD.table(exp.3.nearest,0.10,variables)
tabNEAREST.3
dim(exp.3.nearest)
table(exp.3.nearest$treatment)

exp.3.maha<-mahalanobis.match(dati.3,"treatment",tomatch.3,FALSE, 0.8)
tabMAHALANOBIS.3<-ASD.table(exp.3.maha, 0.10,variables)
tabMAHALANOBIS.3
dim(exp.3.maha)
table(exp.3.maha$treatment)


#EXP5
exp.5.nearest<-nearest_neighbor(dati.5,"treatment",tomatch.5,FALSE, 0.02)
tabNEAREST.5<-ASD.table(exp.5.nearest,0.10,variables)
tabNEAREST.5
dim(exp.5.nearest)
table(exp.5.nearest$treatment)

exp.5.maha<-mahalanobis.match(dati.5,"treatment",c(potential.5),FALSE, 0.02)
tabMAHALANOBIS.5<-ASD.table(exp.5.maha, 0.10,variables)
tabMAHALANOBIS.5
dim(exp.5.maha)
table(exp.5.maha$treatment)


#EXP7
```

```
exp.7.nearest<-nearest_neighbor(dati.7,"treatment",tomatch.7,FALSE, 0.02)
tabNEAREST.7<-ASD.table(exp.7.nearest,0.10,variables)
tabNEAREST.7
dim(exp.7.nearest)
table(exp.7.nearest$treatment)

exp.7.maha<-mahalanobis.match(dati.7,"treatment",tomatch.7,FALSE, 0.06)
tabMAHALANOBIS.7<-ASD.table(exp.7.maha, 0.10,variables)
tabMAHALANOBIS.7
dim(exp.7.maha)
table(exp.7.maha$treatment)



#EXP9
dati.9$treatment<-Invert.treatment(dati.9)
exp.9.nearest<-nearest_neighbor(dati.9,"treatment",potential.9,FALSE, 0.1)
tabNEAREST.9<-ASD.table(exp.9.nearest,0.10,variables)
tabNEAREST.9
dim(exp.9.nearest)
table(exp.9.nearest$treatment)

exp.9.maha<-mahalanobis.match(dati.9,"treatment",potential.9,FALSE, 1.5)
tabMAHALANOBIS.9<-ASD.table(exp.9.maha, 0.10,variables)
tabMAHALANOBIS.9
dim(exp.9.maha)
table(exp.9.maha$treatment)


#MATCHED SAMPLES

nearest.1<-exp.1.nearest
nearest.3<-exp.3.nearest
nearest.5<-exp.5.nearest
nearest.7<-exp.7.nearest
nearest.9<-exp.9.nearest


maha.1<-exp.1.maha
maha.3<-exp.3.maha
maha.5<-exp.5.maha
maha.7<-exp.7.maha
maha.9<-exp.9.maha


################################################################################
#################################OUTCAME PHASE##################################
################################################################################



#FUNCTION
NB_regression<-function(dat, index.y, index.x){
  library(MASS)
  if(is.numeric(index.x)){
    variables<-colnames(dat)[index.x] }
  else{variables<-index.x}
  if(is.numeric(index.y)){
    outcome<-colnames(dat)[index.y]}
  else{outcome<-index.y}
  frml<-as.formula(paste(outcome, paste(c(variables), sep=" ", collapse="
+"),sep="~"))
  options(digits = 5)
  Negative_binomial<-glm.nb(frml, dat)
  summary_NB<-summary(Negative_binomial)
  coef<-Negative_binomial$coeff[[2]]
  Exponential.coefficients_NB<-exp(Negative_binomial$coeff)
```

```
  Exponential.CI_NB<-exp(confint(Negative_binomial))
  results<-c("NEGATIVE
BINOMIAL"=list("Model"=Negative_binomial,"Summary"=summary_NB,
                                        "AvgPM"=coef,
                                        "Exponential of
coefficients"=Exponential.coefficients_NB,
                                        "Exponential of confidence
interval"=Exponential.CI_NB))
  return(results)
}
adjust<-variables[-c(18,19,30)]

#ONE SAMPLE
M.OS<-NB_regression(one.sample,"Tot_num.All.cause.admissions",
c("avgPM","offset(log(Total_den))"))
M.OS

M.OS.ADA<-NB_regression(one.sample,"Tot_num.All.cause.admissions",
c("avgPM",adjust,"offset(log(Total_den))"))
M.OS.ADA

M.OS.NEAR<-NB_regression(one.sample.near,"Tot_num.All.cause.admissions",
c("avgPM","offset(log(Total_den))"))
M.OS.NEAR

M.OS.NEAR.ADA<-NB_regression(one.sample.near,"Tot_num.All.cause.admissions",
c("avgPM",adjust,"offset(log(Total_den))"))
M.OS.NEAR.ADA

M.OS.MAHA<-NB_regression(one.sample.maha,"Tot_num.All.cause.admissions",
c("avgPM","offset(log(Total_den))"))
M.OS.MAHA

M.OS.MAHA.ADA<-NB_regression(one.sample.maha,"Tot_num.All.cause.admissions",
c("avgPM",adjust,"offset(log(Total_den))"))
M.OS.MAHA.ADA


#5 EXPERIMENTS
M.EXP<-lapply(list(dati.1,dati.3,dati.5,dati.7,dati.9),
function(x) NB_regression(x,"Tot_num.All.cause.admissions" ,
c("avgPM","offset(log(Total_den))")))
M.EXP[[1]]
M.EXP[[2]]
M.EXP[[3]]
M.EXP[[4]]
M.EXP[[5]]

dim(one.sample)

lapply(list(one.sample.near, one.sample.maha), function(x) dim(x))
lapply(list(dati.1,dati.3,dati.5,dati.7,dati.9),
       function(x) dim(x))

lapply(list(nearest.1,nearest.3,
            nearest.5,nearest.7,nearest.9,
            maha.1,maha.3,maha.5,maha.7, maha.9),
       function(x) dim(x))

lapply(list(NEAR.EXP, MAHA.EXP),
       function(x) dim(x))

M.EXP.ADA<-lapply(list(dati.1, dati.3,dati.5,dati.7,dati.9),
function(x)NB_regression(x,"Tot_num.All.cause.admissions",c("avgPM",adjust,"o
ffset(log(Total_den))")))
M.EXP.ADA[[1]]
M.EXP.ADA[[2]]
```

```
M.EXP.ADA[[3]]
M.EXP.ADA[[4]]
M.EXP.ADA[[5]]


M.EXP.M<-lapply(list(nearest.1,nearest.3,
nearest.5,nearest.7,nearest.9,
maha.1,maha.3,maha.5,maha.7, maha.9),function(x)NB_regression(x,
"Tot_num.All.cause.admissions", c("avgPM","offset(log(Total_den))")))



lapply(list(nearest.1,nearest.3,
            nearest.5,nearest.7,nearest.9,
            maha.1,maha.3,maha.5,maha.7, maha.9),
       function(x) range(x$avgPM))

#NEAR
M.EXP.M[[1]]
M.EXP.M[[2]]
M.EXP.M[[3]]
M.EXP.M[[4]]
M.EXP.M[[5]]

#NEAR
M.EXP.M[[6]]
M.EXP.M[[7]]
M.EXP.M[[8]]
M.EXP.M[[9]]
M.EXP.M[[10]]

M.EXP.M.ADA<-lapply(list(nearest.1,nearest.3,
nearest.5,nearest.7,nearest.9,
maha.1,maha.3,maha.5,maha.7, maha.9),
function(x)NB_regression(x, "Tot_num.All.cause.admissions",
c("avgPM",adjust,"offset(log(Total_den))")))

#NEAR
M.EXP.M.ADA[[1]]
M.EXP.M.ADA[[2]]
M.EXP.M.ADA[[3]]
M.EXP.M.ADA[[4]]
M.EXP.M.ADA[[5]]

#MAHA
M.EXP.M.ADA[[6]]
M.EXP.M.ADA[[7]]
M.EXP.M.ADA[[8]]
M.EXP.M.ADA[[9]]
M.EXP.M.ADA[[10]]


NEAR.EXP<-rbind(nearest.1,nearest.3,
nearest.5,nearest.7,nearest.9)


MAHA.EXP<-rbind(maha.1,maha.3,maha.5,
maha.7,maha.9)

M.EXP.NEAR<-NB_regression(NEAR.EXP,"Tot_num.All.cause.admissions",
c("avgPM","offset(log(Total_den))"))
M.EXP.NEAR

M.EXP.NEAR.ADA<-NB_regression(NEAR.EXP,"Tot_num.All.cause.admissions",
c("avgPM",adjust,"offset(log(Total_den))"))
M.EXP.NEAR.ADA
```

```
M.EXP.MAHA<-NB_regression(MAHA.EXP,"Tot_num.All.cause.admissions",
c("avgPM","offset(log(Total_den))"))
M.EXP.MAHA

M.EXP.MAHA.ADA<-NB_regression(MAHA.EXP,"Tot_num.All.cause.admissions",
c("avgPM",adjust,"offset(log(Total_den))"))
M.EXP.MAHA.ADA



####SPLINES
library(splines)
#SAMPLE MATCHED
#sample no obs >15 of pm2.5
ORIGINAL<-one.sample[-which(one.sample$avgPM>12),]
summary(ORIGINAL)
ORIGINAL<-ORIGINAL[order(ORIGINAL$avgPM),]
head(ORIGINAL)

NEAR.EXP.SPLINE<-NEAR.EXP[-which(NEAR.EXP$avgPM>12),]
summary(NEAR.EXP.SPLINE$avgPM)
quantile(NEAR.EXP.SPLINE$avgPM, c(0.05,0.25,0.75,0.95))

MAHA.EXP.SPLINE<-MAHA.EXP[-which(MAHA.EXP$avgPM>12),]
summary(MAHA.EXP.SPLINE$avgPM)
quantile(MAHA.EXP.SPLINE$avgPM, c(0.05,0.25,0.75,0.95))

M.EXP.NEAR_spline.12<-
NB_regression(NEAR.EXP.SPLINE,"Tot_num.All.cause.admissions",
c("ns(avgPM,knots=c(6.35,7.97,9.70,11.19))","offset(log(Total_den))"))
M.EXP.NEAR_spline.12$`NEGATIVE BINOMIAL.Summary`

M.EXP.MAHA_spline.12<-
NB_regression(MAHA.EXP.SPLINE,"Tot_num.All.cause.admissions",
c("ns(avgPM,knots=c(6.21,7.87,9.57,10.84))","offset(log(Total_den))"))
M.EXP.MAHA_spline.12$`NEGATIVE BINOMIAL.Summary`


PREDICT.DT<-data.frame(avgPM=ORIGINAL$avgPM,Total_den=1)
PREDICT.NEAR<-cbind(ORIGINAL, predict(M.EXP.NEAR_spline.12$`NEGATIVE
BINOMIAL.Model`,newdata=PREDICT.DT, type="response", se.fit=T))
tail(PREDICT.NEAR)

PREDICT.MAHA<-cbind(ORIGINAL, predict(M.EXP.MAHA_spline.12$`NEGATIVE
BINOMIAL.Model`,newdata=PREDICT.DT, type="response", se.fit=T))
tail(PREDICT.MAHA)

names(PREDICT.MAHA)
PREDICT.NEAR$Method<-"Nearest-neighbor"
PREDICT.MAHA$Method<-"Mahalanobis distance"

PREDICT.SPLINE.12<-rbind(PREDICT.NEAR,PREDICT.MAHA)

library(ggplot2)
ggplot(data=PREDICT.SPLINE.12, aes(x=avgPM,
y=(Tot_num.All.cause.admissions/Total_den)*1000, group=Method))+
  geom_point(alpha=0.05) +
  geom_line(aes(x=avgPM, y=fit*1000, colour=Method))+
  #geom_smooth(method="auto",se=TRUE)+
  theme_bw() +
  ylim(100,325)+
  scale_x_continuous(breaks=round(seq(min(PREDICT.SPLINE.12$avgPM),
max(12.5),by=0.5),1))+
  #geom_vline(xintercept=c(7.83,8.65,9.36),linetype="dashed") +
  ggtitle(NULL) +
  theme( plot.title=element_text(vjust=1.0) ) +
```

```
  xlab(expression(paste("PM"["2.5 "],mu,g,"/",m^{3}))) +
  theme( axis.title.x = element_text(vjust=-.5) ) +
  ylab("Predicted all-cause hospital admissions rate (per 1000 person-year)")
+
  theme( axis.title.y = element_text(vjust=1.0) ,
         legend.position = "bottom")
```