Human Exposure Factors as Potential Determinants of the Heterogeneity in City-Specific Associations between PM$_{2.5}$ and Mortality

Lisa K. Baxter, ScD[1,*], **Kathie Dionisio, ScD**[2], **Prachi Pradeep, PhD**[3,4], **Kristen Rappazzo, PhD**[1], and **Lucas Neas, ScD**[1]

[1]National Health and Environmental Effects Research Laboratory, Office of Research and Development, United States Environmental Protection Agency

[2]National Exposure Research Laboratory, Office of Research and Development, United States Environmental Protection Agency

[3]National Center for Computational Toxicology, Office of Research and Development, United States Environmental Protection Agency

[4]Oak Ridge Institute for Science and Education, Oak Ridge, Tennessee

Abstract

Multi-city population-based epidemiological studies of short-term fine particulate matter (PM$_{2.5}$) exposures and mortality have observed heterogeneity in risk estimates between cities. Factors affecting exposures, such as pollutant infiltration, which are not captured by central-site monitoring data, can differ between communities potentially explaining some of this heterogeneity.

This analysis evaluates exposure factors as potential determinants of the heterogeneity in 312 Core-Based Statistical Areas (CBSA) -specific associations between PM$_{2.5}$ and mortality using inverse variance weighted linear regression. Exposure factor variables were created based on data on housing characteristics, commuting patterns, heating fuel usage, and climatic factors from national surveys. When survey data were not available, air conditioning (AC) prevalence was predicted utilizing machine learning techniques.

Across all CBSAs, there was a 0.95% (Interquartile range (IQR) of 2.25) increase in non-accidental mortality per 10 μg/m^3 increase in PM$_{2.5}$ and significant heterogeneity between CBSAs. CBSAs with larger homes, more heating degree days, a higher percentage of home heating with oil had significantly (p < 0.05) higher health effect estimates, while cities with more gas heating had significantly lower health effect estimates. While univariate models did not explain much of heterogeneity in health effect estimates (R^2 < 1%), multivariate models began to explain some of the observed heterogeneity (R^2 = 13%).

*corresponding author Lisa Baxter, 109 T.W. Alexander Drive, MD B305-02, Research Triangle Park, NC 27711, Office phone: 919-541-0672, Fax: 919-541-1440, baxter.lisa@epa.gov.

Conflict of Interest

The authors declare there are no competing financial interests in relation to the work described.

Keywords

fine particulate matter; exposure factors; epidemiology; meta-regression

Introduction

Multi-city population-based epidemiological studies of short-term fine particulate matter ($PM_{2.5}$) exposures and mortality have provided evidence of heterogeneity in risk estimates between communities and cities (1, 2). This heterogeneity remains a key source of uncertainty in the examination of the relationship between short-term $PM_{2.5}$ exposures and mortality. These differences in PM mortality risk estimates has often been attributed to exposure measurement error (3), differences in particle composition (4) , and/or city-specific differences such as demographics (5). $PM_{2.5}$ concentrations from central-site monitors, such as those reported in the United States Environmental Protection Agency's Air Quality System, are often used as a surrogate for exposure in epidemiological studies. Use of central-site monitoring data has been shown to introduce both a Berkson error component, a result of using aggregated instead of personal exposure data, and a classical error component, a result of the difference between the aggregated exposure data and the true ambient $PM_{2.5}$ concentrations (6). Berkson error would not bias the health effect estimates, but would lead to an increased variance, while classical error can lead to bias (6, 7).

If the relationship between central-site monitor measurements and estimates of personal exposure to ambient concentrations varies by city these differences could explain heterogeneity in $PM_{2.5}$ -mortality estimates. Studies focusing on PM_{10} have examined the contributions of exposure factors as modifiers to the association between exposure and mortality, with both studies observing differences in mortality effects estimates related to differences in indoor exposures to PM_{10} of outdoor origin (8, 9). While the infiltration characteristics differ between PM_{10} and $PM_{2.5}$, higher air pollution risks have been reported for cities with higher overall air exchange rates (AERs) or pollutant infiltration efficiencies (10-13). A number of factors related to home characteristics can influence the infiltration of ambient air into the home, including age of home (14, 15) and housing type (i.e., multi- vs. single-family home) (16, 17). Previous studies have also observed the prevalence of central air conditioning as a predictor of heterogeneity in the association between $PM_{2.5}$ and mortality (12, 18). Average outside temperatures can affect infiltration by influencing occupant behaviors such as air conditioning use and opening of windows (19). These factors may not only be related to infiltration but may also be surrogates for socio-economic factors.

Meteorology will also impact heating use and type of heating fuel. The type of heating fuel used can affect the composition of the PM2.5 outside of the home. For example, higher concentrations of nickel and vanadium have be associated with residual oil-burning (20) and studies have suggested that PM2.5 effects were stronger in locations with higher residual oil-burning (21, 22). The amount of time spent in-vehicle can affect the amount and composition of particulate matter a person is exposed it. In-vehicle air pollution measurement studies have indicated that concentrations of pollutants inside cars and buses are considerably higher than those recorded at nearby ambient monitors (23). Exposure

models suggest that even a small amount of time spent in vehicles may contribute significantly to the average daily personal PM exposure (24).

The objective of this analysis is to evaluate potential exposure factors as determinants of heterogeneity in city-specific associations between $PM_{2.5}$ and mortality. These associations were developed for 312 Core Based Statistical Areas (CBSAs) and Metropolitan Divisions (MD) across the US from 1999–2005. Factors investigated include variables related to home age, type and size, commuting times and distances, prevalence of AC in the CBSA, type of heating fuel used, and annual heating and cooling degree days.

Methods

The association between daily $PM_{2.5}$ concentrations and non-accidental mortality was established for 312 CBSAs and MDs across the continental United States for the years 1999–2005 using Poisson time-series models. Exposure factor variables were created based on data on housing characteristics, commuting patterns, heating fuel usage, and climatic factors from national surveys. Survey data was not available on AC prevalence for all CBSAs therefore this factor was predicted utilizing machine learning techniques for CBSA lacking data. Meta-regression, using inverse variance weighted linear regression, was then applied to the log rate ratios to determine the influence of these exposure factors on the observed heterogeneity in $PM_{2.5}$-mortality effect estimates.

Health Effect Estimates

Area-specific associations between daily counts of total non-accidental mortality and daily ambient concentrations of $PM_{2.5}$ at lag 1 for 1999 through 2005 were calculated using time series methods adjusting for time/season, day of week, current temperature and dew point temperature, and individual lagged temperature at lags 1–3. For each CBSA/MD, an estimate of the area-specific association (effect estimate) and its standard error was obtained. All analyses were conducted with area-specific inverse-variance weights, that is, the inverse of the variance.

Data were spatially aggregated into all 929 Core-based Statistical Areas (CBSA) and their 31 Metropolitan Divisions (MD) as defined by the White House Office of Management and Budget in February, 2013(25). The 31 MDs into which 11 CBSAs were divided were considered as distinct areas and will henceforth simply be described as CBSAs. These 949 unique, multi-county areas include all urban clusters with a population of at least 10,000 and comprise about 94% of the U.S. population in 2010. Every attempt was made to retain each CBSA/MD that had at least 60 days of meteorological and air quality data, but not every CBSA/MD had air quality monitors or weather stations.

Detailed cause-specific mortality data for 1999–2005 with individual-level data on county of residence and date of death were obtained from the US National Center for Health Statistics and converted to daily counts of non-accidental mortality (ICD 10 excluding S, T, U, V, W, X, Y, Z) by county of residence and by major cause of death (http://www.cdc.gov/nchs/about.htm). Meteorological data for 1985–2005, hourly temperature and dew point

temperature along with the latitude and longitude of each weather station, were obtained from the US National Climatic Data Center (26).

All available air quality data from population-based air quality monitors on a daily 24-hour mean for fine particulate matter ($PM_{2.5}$) were obtained from the US Environmental Protection Agency's Air Quality System (27). Air quality monitoring was available only from 1999 – 2005 for $PM_{2.5}$. Across monitor averages when multiple monitors sites are present in a city has been shown to lead to less bias in time-series studies (7). Less bias is also expected with pollutants such as $PM_{2.5}$ that are spatially homogeneous (28). Multiple air quality monitors within a CBSA/MD were therefore summarized as described in Baxter et al. 2017 (29) where daily values were standardized using all values within a given CBSA/MD instead of using all values for its corresponding monitor.

For this analysis, the natural logarithm of the conditional mean daily mortality count (Y) was modeled as a linear function of a selected air pollutant (AP) adjusting for covariates (X) including a natural spline for time (*Time*) with seven degrees of freedom per year, a one degree of freedom per year interaction of time with binary indicators of age group (G), binary indicators for day of week (DOW), current day's (lag 0) temperature ($Temp_0$) and dew point temperature (Dew_0), and separate natural spline terms with three degrees of freedom for each lag of temperature ($Temp_{-m}$) between 1 and 3:

$$
\ln(E[Y \mid X]) = \alpha + \beta_{AP}AP + ns(Time, 7\,df\,/year) + \sum_{\kappa=1}^{2} G_{\kappa} * ns_{\kappa}(Time, 1df\,/year)
$$
$$
+ \sum_{i=1}^{6} \beta_i DOW_i + ns(Temp_0, 3\,df) + ns(Dew_0, 3\,df)
$$
$$
+ \sum_{m=1}^{3} ns_m(Temp_{-m}, 3\,df)
$$

The time-series data were organized into three age groups: 0 to 64 years of age at death, 65– 74 years and 75 years and older. By modeling the lagged effects of temperature with separate smooths at each lag (1-3) the prior constraints on individual lags were relaxed. The effect estimates for each CBSA/MD are expressed as a percent change in nonaccidental mortality for a 10 μg/m^3 increment in the 24-hour mean of $PM_{2.5}$ at lag 1. Based on data completeness health effect estimates were generated for 312 CBSAs.

Exposure Factors

Exposure factors related to housing characteristics, household heating, commuting to work, and meteorology were constructed from 2000 U.S. Census data (30). The information was downloaded at the county level and aggregated to the CBSA-level using weighted averages. Depending on the variable the were either counts of population or housing units. Meteorological factors (cooling degree days, heating degree days, and average annual relative humidity) were constructed using AQS data (27). Each of the 312 CBSA was assigned a single value per exposure factor. Additional details on the construction of exposure factor variables can be found in supplemental material.

The prevalence of central air conditioning is also a potential predictor of interest. However, of the 312 CBSAs included in the analysis, only 54 included data on air conditioning prevalence from the 2011 American Housing Survey (31). Therefore, machine learning techniques were used to predict prevalence of central AC for the 258 CBSAs where AC data were not available. Variables known to be predictive of central AC prevalence (3, 32) were chosen a priori (Table 1). Along with the previously described exposure factors, additional meteorological and poverty related variables were included as potential predictors of central AC prevalence. Construction of the annual average relative humidity and percent of families below poverty level variables are described in supplemental information. These factors were used to train and test lasso regression, support vector machine (SVM), and random forest models to predict prevalence of central AC by CBSA. The specific machine learning models used were methods amendable to smaller datasets. Previous publications have also used machine learning methods with similarly small sample sizes (33-35). Machine learning models were implemented in R (version 3.1.2) using the 'caret' package (http://topepo.github.io/caret/index.html) available on CRAN (https://cran.r-project.org/web/packages/caret/).

The models were evaluated using 4-fold external cross-validation such that the 75% data were used for training the models and 25% data were used for testing. Each training set was further evaluated using 10-fold internal cross-validation such that 90% of the data were used for training and 10% data were used for testing. For each model type, limits of 0 to 1 were set on the predicted values for AC prevalence. For the random forest models, individual models were tested using 100, 500, and 1000 trees. The hyper-parameters of each model were tuned using 10-fold CV resampling.

Meta-regression

The heterogeneity in 312 city-specific $PM_{2.5}$-mortality health effect estimates were examined using a fixed-effects inverse variance weighted linear regression. First, univariate meta-regression was performed by regressing the individual exposure factors on the outcome measure. There was also an interest in developing multivariate models to examine several exposure factors together. Given the similarities between some of the exposure factors it was anticipated that some would be highly correlated with one another. Therefore, correlations between all factors were calculated. For those factors strongly correlated (greater than 0.7 or less than −0.7), the factor with the highest F-statistic from the univariate models was included as a potential predictor in the multivariate models. A final multivariate model was determined using backwards selection. Under this approach, the model was first fitted with all potential predictors. The model was run iteratively with the least significant variable dropped at each iteration of the model, until all covariates were significant at the chosen critical level ($p < 0.05$).

Results

Health Effect Estimates

Overall there was a 0.95% (Interquartile range (IQR) of 2.25) increase in non-accidental mortality per 10 μg/m3 increase in $PM_{2.5}$. The CBSA/MD-specific health effect estimates

are shown on the map in Figure 1. The geographic variation shows a clustering of positive associations along the northeastern corridor between Baltimore, MD and Augusta, ME with a scattering of additional areas with strong positive associations. Significant heterogeneity was observed among the CBSA-specific estimates with a Q-statistic p-value < 0.0001

Summary Statistics of Exposure Factor Variables

The medians, interquartile ranges, and coefficients of variation (CV) across CBSAs, except for central air conditioning, are presented in (Table 1). The variables are grouped into 5 different categories: housing characteristics, commuting, household heating, metrological factors, and poverty measures.

Predictions of central AC prevalence

Supplemental Table 1 presents the exposure factors selected as potential predictors of air conditioning prevalence. These include median home age, median number of rooms in the residence, cooling degree days (CDD), average annual relative humidity, and percentage of families below poverty level.

Mean and range for root mean squared error (RMSE) and R^2 values across the four mutually exclusive external cross-validation testing groups of CBSAs with known AC are given in Supplemental Table 2. Mean RMSE for both internal and external cross-validation across models ranged from 0.11–0.12, a margin of error consistent with known error introduced when collecting data on prevalence of AC through a survey. Mean R^2 across the three methods tested (lasso, linear SVM, random forest) from 0.82–0.84 and 0.75–0.78 for internal and external cross-validation, respectively. Results of 10-fold internal cross-validation for random forest models with 100, 500, and 1000 trees showed no difference (results not shown), thus a random forest model with n=100 trees were used for subsequent analysis. As no one model outperformed the others, a consensus model was chosen which averaged prediction results from the lasso, linear SVM, and random forest (100 trees) models.

The final tuning parameters used in the models run on the full data set, which are used to obtain the consensus model predictions are: fraction = 0.9 (lasso), cost = 1 and gamma = 2 (SVM linear), mtry = 3 (random forest – 100 trees; mtry represents the number of variables randomly sampled as candidates at each split). Further details are provided in Supplemental Information and in Supplemental Table 3. Predictions of central AC prevalence from the consensus method were used in subsequent meta-regressions as predictors of heterogeneity in the association between $PM_{2.5}$ and mortality.

Comparison of the range and distribution of values for predictors between the CBSAs with known central AC prevalence and those with unknown central AC prevalence are shown in Table 2. The median percentage of central AC in the empirical dataset (known from the survey data) was higher than the predicted dataset. The difference can be attributed to the greater number of cooling degree days in the empirical dataset compared to the predicted dataset (Table S1). The geographical patterns of the prevalence of central AC for the empirical dataset are similar to AC predictions (Figure 2) with a higher prevalence of AC in the southeast and southwest compared to other regions of the country.

Correlations between exposure factors

The correlations between the exposure factors are presented in Figure 3 and in Supplemental Table 4S (Abbreviation Key in Supplemental Information). Several of the exposure factors were strongly correlated with one another ($|r| > 0.7$). These included median number rooms total and owner occupied; all single family homes and detached single family homes; homes with more than 2 units and homes with more than 5 units; median number of rooms (renter occupied) and homes with more than 5 units; commuting alone and using public transportation; using electric heat and gas heat; heating degree days and homes with electric heat; and prevalence of air conditioning and cooling degree days. For each pair listed, the factors with the higher F-statistic from the univariate models (Table 4) was included as a potential predictor in the multivariate regression, with one exception. The homes with more than 5 units ('Homes-5+') variable was strongly correlated with both the multi-family homes variable and the median number of rooms in renter occupied homes variable. In the former instance the Homes-5+ variable had the lower F statistic, in the latter instance, the Homes-5+ variable has the higher F statistic. Because in the latter instance the F statistics for the two correlated variables were close in magnitude, the decision was made to drop the Homes-5+ variable which took care of the high correlation for both pairs.

Meta-regression results

A second stage analysis (meta-regression) was conducted in which the individual exposure factors were regressed against the outcome measure for each individual CBSA weighted by their corresponding inverse variances. The assumption of linearity was examined using spline models and no substantial departure of linearity was observed (data not shown). Table 3 presents the coefficients, 95% confidence interval, F-statistic, and p-values of the univariate regressions. Significant predictors ($p < 0.05$) for the variability in health effect estimates include median number of rooms in residences that were owner occupied, percentage of duplex homes, median number of rooms in residence, percentage of multi-family homes, percentage of homes using oil heating, percentage of homes using gas heating, and heating degree days. Larger homes were associated with larger health effect estimates. For example, an increase in the IQR number of rooms (0.5 rooms) was associated with a 0.258% increase in mortality per $10\mu g/m^3$ increase in $PM_{2.5}$. A 37% (the IQR) increase in the percentage of homes with gas heating was associated with a 0.412% decrease in mortality while a 7 % (the IQR) increases in the percentage of homes with oil heating was associated with a 0.148% increase in mortality. Finally, a 0.319% increase in mortality was observed with an increase in heating degree days (IQR = 3477 days). However, the R^2 for all univariate models were less than 1% (results not shown) suggesting that a single variable did not explain much of the heterogeneity. Therefore, multivariate models were developed.

The results of the multivariate regression analysis using the backward selection approach are shown on Table 4. The F-statistics between cooling degree days and prevalence of AC were not very different; therefore, 2 sets of potential predictors were considered, one with cooling degree days (Model 1) and one with central AC (Model 2), and run through the backward selection process. The final models resulting from the backward selection approach using either central AC or cooling degree days both include fraction of duplex homes, fraction of homes with gas heating and median number of rooms in a renter occupied home. The

coefficients for cooling degree days (−0.228%), heating degree days (0.448%), and homes heating with gas (−0.471% in Model 1 and −0.427%) were similar to the univariate results. However, the association of median number of rooms, renter occupied was negative in Model 1 with a 0.123% (per IQR in median number of rooms) decrease in mortality per 10 μg/m^3 increase in PM$_{2.5}$ and positive in Model 2 with a 0.215% increase in mortality. In addition to the aforementioned variables, Model 1 also included median home age and heating degree days.

Discussion

The association between PM$_{2.5}$ and mortality were estimated for 312 CBSAs, with a pooled estimate of 0.95% (Interquartile range (IQR) of 2.25) increase in non-accidental mortality per 10 μg/m^3 increase in PM$_{2.5}$. Significant heterogeneity was observed across the health effect estimates. A meta-regression was performed to examine the potential for a variety of exposure factors to explain the heterogeneity in the effect estimates. In the univariate meta-regression analyses, various exposure factors related to housing size, heating fuel type, and meteorology were significant but did not explain a large portion of the variability. The final multivariate models from the meta-regression analyses included gas heating use, heating degree days, cooling degree days, and variables for home size and age. These models explained 11–13% of the variability in the health effect estimates.

The exposure factors were available from national surveys for all 312 CBSAs, except for prevalence of air conditioning. The prevalence of central AC was only available from the AHS for a limited number of CBSAs (n=54). Machine learning techniques were utilized to predict the prevalence of central AC for CBSAs where data were missing. A consensus model was chosen with an R^2 of 0.78 indicating a strong relationship between the training set (central AC information from the AHS) and the predicted AC variable. This is one of the first studies to employ machine learning techniques in this context and would serve as a baseline for newer improved studies. Previous studies have either used data from the AHS (2, 10), questionnaire (36, 37) to determine the prevalence of air conditioning in a limited number of cities/communities, or extrapolated from known areas with similar housing characteristics and weather to estimate AC prevalence (38). The high R^2 of the model suggests that this approach produced good estimates of AC prevalence.

In the univariate meta-regression analysis, CBSAs with larger homes have higher health effect estimates than CBSAs with smaller homes. Air exchange rates have been found to be higher in larger homes resulting in higher exposures to outdoor PM (39) and associations higher in magnitude. Other significant factors were related to heating in the form of heating fuel type and the number of heating degree days. The greater number of heating degree days and homes heated by oil were associated with higher health effect estimates. Conversely, CBSAs with a higher percentage of homes heated with utility gas had lower health effect estimates. The choice of home heating fuel can alter the composition of the PM$_{2.5}$ with higher usage of oil leading to higher levels of nickel, vanadium, and elemental carbon. Previous studies have observed stronger PM$_{2.5}$ health effects in locations with higher residual oil-burning-related pollution (21, 22). A larger number of cooling degree days were also associated with lower health effect estimates. As seen in the correlations CDD the high

correlation (0.87) with AC. Increased prevalence and use of central AC will result in lower infiltration of outdoor air indoors (38) leading to lower health effect estimates (29).

The R^2s for the univariate models were less than 1% suggesting that a single variable was not responsible for the heterogeneity in the health effect estimates. A combination of variables was therefore examined using multivariate modeling. While these models explained 11%(Model 1) and 13% (Model 2) of the variability some of the results were not as expected. Median number of rooms in renter occupied homes was significant in both models; however, the direction of the associations was different. Larger houses typically have higher infiltration compared to smaller houses, since they contain a greater surface area for leaks to develop (15). Similarly, older homes tend to be leakier allowing for more outdoor air to penetrate indoors (15, 40). This increase in exposure to outdoor air was hypothesized to lead to an increase in health effects however the opposite was observed. Housing characteristics are surrogates rather than direct measurements of infiltration and may also be associated with other characteristics such as socioeconomic status. Socioeconomic factors have been shown to have their own relationships with both exposures to air pollution and health effects (5). Finally, while reductions in infiltration will reduce exposures to $PM_{2.5}$ of ambient origin it will also increase exposures to $PM_{2.5}$ generated from indoor sources. This indoor $PM_{2.5}$ may be independently associated with adverse health effects.

Breathing polluted air has been associated with increased mortality and hospitalizations but the magnitude of these effects can be different in different cities and/or regions. Many factors have been suggested to cause differences in concentration-response relationships of ambient particulate matter. These may include differences in the chemical composition of the particles; differences in the populations (i.e. sociodemographic factors); and differences in exposures caused by the ways buildings are ventilated and people's activity patterns. This study examined the role of exposure differences in the heterogeneity of $PM_{2.5}$ mortality effect estimates across the U.S., finding that factors related to infiltration and home heating fuel were significant effect modifiers to the $PM_{2.5}$-mortality relationship and explained some of the observed heterogeneity. Areas with housing characteristics linked to increased infiltration rates or where particular kind of fuel use may be exposed to higher levels of outdoor pollutants and therefore be more vulnerable. National standards may not protect everyone equally, with some populations affected more than others. Determining potential factors that can cause increases in adverse $PM_{2.5}$ -health effects may help identify at-risk populations and develop actions that can reduce risk.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

References

1. 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;295(10): 1127–34. [PubMed: 16522832]

2. Franklin M, Zeka A, Schwartz J. Association between $PM_{2.5}$ and all-cause and specific-cause mortality in 27 US communities. J Expo Sci Environ Epidemiol. 2007;17(3):279–87. [PubMed: 17006435]

3. Baxter L, Franklin M, Özkaynak H, Schultz B, Neas L. The use of improved exposure factors in the interpretation of fine particulate matter epidemiological results. Air Qual Atmos Health. 2013;6(1): 195–204.

4. Franklin M, Koutrakis P, Schwartz J. The role of particle composition and the association between $PM_{2.5}$ and mortality. Epidemiology. 2008;19(5):680–9. [PubMed: 18714438]

5. Bell ML, Ebisu K. Environmental inequality in exposures to airborne particulate matter components in the United States. Environ Health Perspect. 2012;120(12):1699–704. [PubMed: 22889745]

6. Zeger SL, Thomas D, Dominici F, Samet JM, Schwartz J, Dockery D, et al. Exposure measurement error in time-series studies of air pollution: concepts and consequences. Environ Health Perspect. 2000;108(5):419–26. [PubMed: 10811568]

7. Goldman GT, Mulholland JA, Russell AG, Gass K, Strickland MJ, Tolbert PE. Characterization of ambient air pollution measurement error in a time-series health study using a geostatistical simulation approach. Atmos Environ. 2012;57(0):101–8.

8. Chen C, Zhao B, Weschler CJ. Indoor exposure to "outdoor PM10": assessing its influence on the relationship between PM10 and short-term mortality in U.S. cities. Epidemiology. 2012;23(6):870–8. [PubMed: 23018971]

9. Zhou B, Zhao B, Guo X, Chen R, Kan H. Investigating the geographical heterogeneity in PM10-mortality associations in the China Air Pollution and Health Effects Study (CAPES): A potential role of indoor exposure to PM10 of outdoor origin. Atmos Environ. 2013;75:271–23.

10. Bell ML, Dominici F. Effect modification by community characteristics on the short-term effects of ozone exposure and mortality in 98 US communities. Am J Epidemiol. 2008;167(8):986–97. [PubMed: 18303005]

11. Hodas N, Lunden MM, Meng QY, Baxter LK, Özkaynak H, Burke J, et al. Heterogeneity in the fraction of ambient $PM_{2.5}$ found indoors contributes exposure error and may contribute to spatial and temporal differences in reported $PM_{2.5}$ health effect estimates. J Expo Sci Environ Epidemiol. 2012;22(5):448–54. [PubMed: 22617722]

12. Janssen NAH, Schwartz J, Zanobetti A, Suh HH. Air conditioning and source specific particles as modifiers of the effect of PM_{10} on hospital admissions for heart and lung disease. Environ Health Perspect. 2002;110:43–9.

13. Medina-Ramon M, Zanobetti A, Schwartz J. The Effect of Ozone and PM10 on Hospital Admissions for Pneumonia and Chronic Obstructive Pulmonary Disease: A National Multicity Study. Am J Epidemiol. 2006;163(6):579–88. [PubMed: 16443803]

14. Allen R, Larson T, Sheppard L, Wallace L, Liu LJS. Use of real-time light scattering data to estimate the contribution of infiltrated and indoor-generated particles to indoor air. Environ Sci Tech. 2003;37(16):3484–92.

15. Chan WR, Nazaroff WW, Price PN, Sohn MD, Gadgil AJ. Analyzing a database of residential air leakage in the United States. Atmos Environ. 2005;39:3444–55.

16. Koenig JQ, Mar TF, Allen RW, Jansen K, Lumley T, Sullivan JH, et al. Pulmonary effects of indoor- and outdoor-generated particles in children with asthma. Environ Health Perspect. 2005;113(4):499–503. [PubMed: 15811822]

17. Pandian MD, Ott WR, Behar JV. Residential air exchange rates for use in indoor air and exposure modeling studies. J Exp Anal Environ Epidemiol. 1993;3(4):407–16.

18. Bell ML, Ebisu K, Peng RD, Dominici F. Adverse health effects of particulate air pollution: modification by air conditioning. Epidemiology. 2009;20(5).

19. Johnson T, Long T. Determining the frequency of open windows in residences: a pilot study in Durham, North Carolina during varying temperature conditions. J Exp Anal Environ Epidemiol. 2005;15(4):329–49.

20. Ito K, Xue N, Thurston G. Spatial variation of $PM_{2.5}$ chemical species and source-apportioned mass concentrations in New York City. Atmos Environ. 2004;38(31):5269–82.

21. Ito K, Mathes R, Ross Z, Nádas A, Thurston G, Matte T. Fine Particulate Matter Constituents Associated with Cardiovascular Hospitalizations and Mortality in New York City. Environ Health Perspect. 2011;119(4).

22. Bell ML, Ebisu K, Peng RD, Samet JM, Dominici F. Hospital Admissions and Chemical Composition of Fine Particle Air Pollution. Am J Respir Crit Care Med. 2009;179(12):1115–20. [PubMed: 19299499]

23. Weichenthal S, Van Ryswyk K, Kulka R, Sun L, Wallace L, Joseph L. In-Vehicle Exposures to Particulate Air Pollution in Canadian Metropolitan Areas: The Urban Transportation Exposure Study. Environmental Science & Technology. 2014;29(1):597–605.

24. Liu X, Frey HC. Modeling of in-vehicle human exposure to ambient fine particulate matter. Atmos Environ. 2011;45(27):4745–52.

25. White House Office of Management and Budget. Revised Delineations of Metropolitan Statistical Areas, Micropolitan Statistical Areas, and Combined Statistical Areas, and Guidance on Uses ofthe Delineations ofThese Areas. 2013.

26. NCDC. National Oceanic and Atmospheric Association, National Climatic Data Center accessed January 2006 [Available from: http://www.ncdc.noaa.gov/oa/ncdc.html.

27. United States Environmental Protection Agency. United States Environmental Protection Agency, Air Quality System Data Mart accessed October 2012-August 2013 [Available from: http://www.epa.gov/ttn/airs/aqsdatamart.

28. Sarnat SE, Klein M, Sarnat JA, Flanders WD, Waller LA, Mulholland JA. An examination of exposure measurement error from air pollutant spatial variability in time-series studies. J Expo Sci Environ Epidemiol. 2010;20:135–46. [PubMed: 19277071]

29. Baxter LK, Crooks JL, Sacks JD. Influence of exposure differences on city-to-city heterogeneity in $PM2_{.5}$-mortality associations in US cities. Environmental Health. 2017;16(1):1–8. [PubMed: 28049482]

30. United States Census Bureau. United States Census Bureau, American FactFinder 2005 [Available from: http://factfinder.census.gov/faces/nav/jsf/pages/programs.xhtml?program=acs.

31. American Housing Survey (2011) U.S. Census Bureau. American Housing Survey for the United States. Washington, DC; 2009.

32. Baxter LK, Sacks JD. Clustering cities with similar fine particulate matter exposure characteristics based on residential infiltration and in-vehicle commuting factors. Sci Total Environ. 2014;470–471(0):631–8.

33. Prachi P, Povinelli RJ, Merrill SJ, Bozdag S, Sem DS. Novel Uses of In Vitro Data to Develop Quantitative Biological Activity Relationship Models for in Vivo Carcinogenicity Prediction. Molecular Informatics. 2015;34(4):236–45. [PubMed: 27490169]

34. Zhang Y, Dong Z, Phillips P, Wang S, Ji G, Yang J, et al. Detection of subjects and brain regions related to Alzheimer's disease using 3D MRI scans based on eigenbrain and machine learning. Frontiers in Computational Neuroscience. 2015;9(66).

35. Paixão P, Gouveia LF, Morais JAG. Prediction of the in vitro intrinsic clearance determined in suspensions of human hepatocytes by using artificial neural networks. European Journal of Pharmaceutical Sciences. 2010;39(5):310–21. [PubMed: 20056146]

36. Baxter LK, Clougherty JE, Laden F, Levy JI. Predictors of concentrations of nitrogen dioxide, fine particulate matter, and particle constituents inside of lower socioeconomic status urban homes. J Expo Sci Environ Epidemiol. 2007;17(5):433–44. [PubMed: 17051138]

37. Breen MS, Breen M, Williams RW, Schultz BD. Predicting residential air exchange rates from questionnaire and meterology: model evaluation in central North Carolina. Environ Sci Technol. 2010;44(24):9349–56. [PubMed: 21069949]

38. Baxter LK, Stallings C, Smith L, Burke J. Probabilistic estimation of residential air exchange rates for population-based human exposure modeling. J Expo Sci Environ Epidemiol. 2017;27(2):227–34. [PubMed: 27553990]

39. Breen MS, Schultz BD, Sohn MD, Long T, Langstaff J, Williams R, et al. A review of air exchange rate models for air pollution exposure assessments. J Expos Sci Environ Epidemiol. 2014;24(6): 555–63.

40. Persily A, Musser A, Emmerich SJ. Modeled infiltration rate distributions for U.S. housing. Indoor Air. 2010;20(6):473–85. [PubMed: 21070374]
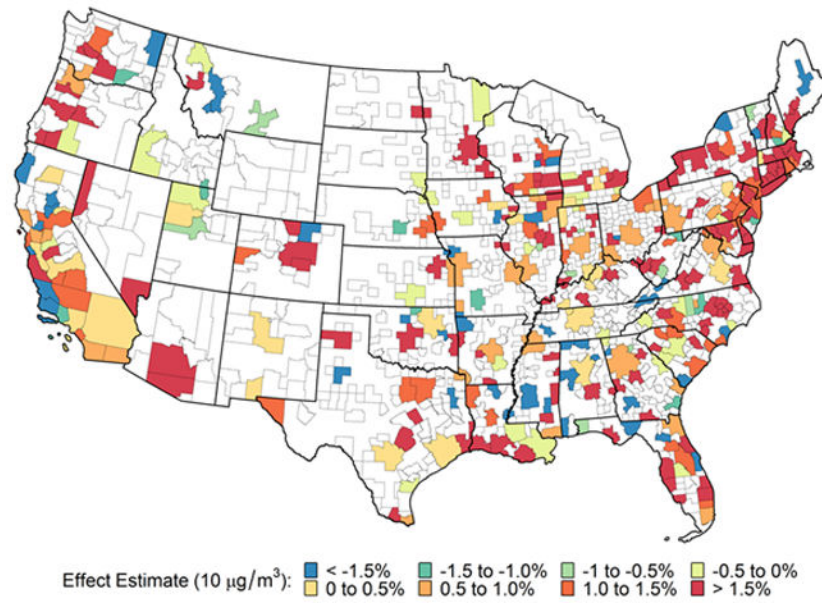
Figure 1.
Area-specific associations of total non-accidental mortality and fine particulate matter (PM2.5) at lag 1: 312 US Core-Based Statistical Areas and Metropolitan divisions

A/C Prevalence by CBSA
297 US Core-Based Statistical Areas and Metropolitan Divisions
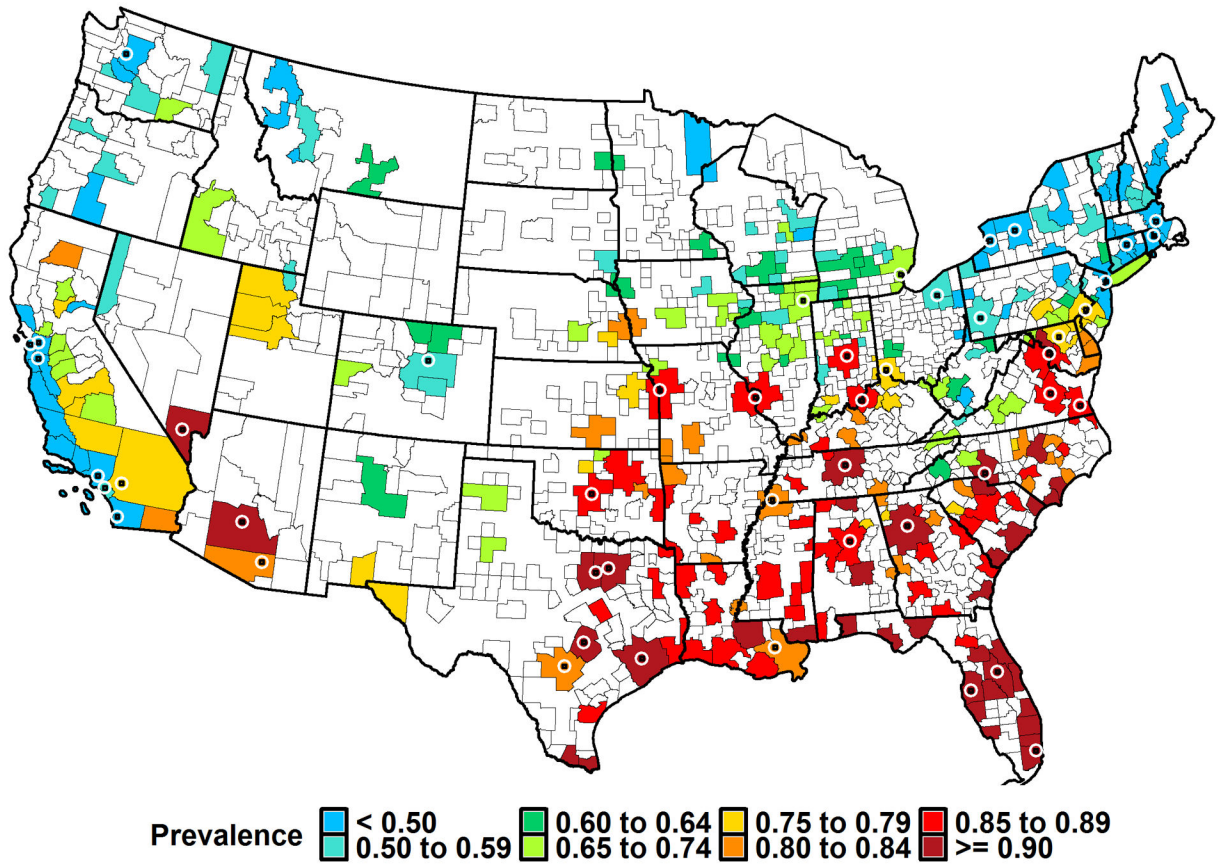

Prevalence | < 0.50 | 0.60 to 0.64 | 0.75 to 0.79 | 0.85 to 0.89
0.50 to 0.59 | 0.65 to 0.74 | 0.80 to 0.84 | >= 0.90

Figure 2:
Prevalence of predicted air conditioning by CBSA (black dots designate CBSAs from training set)
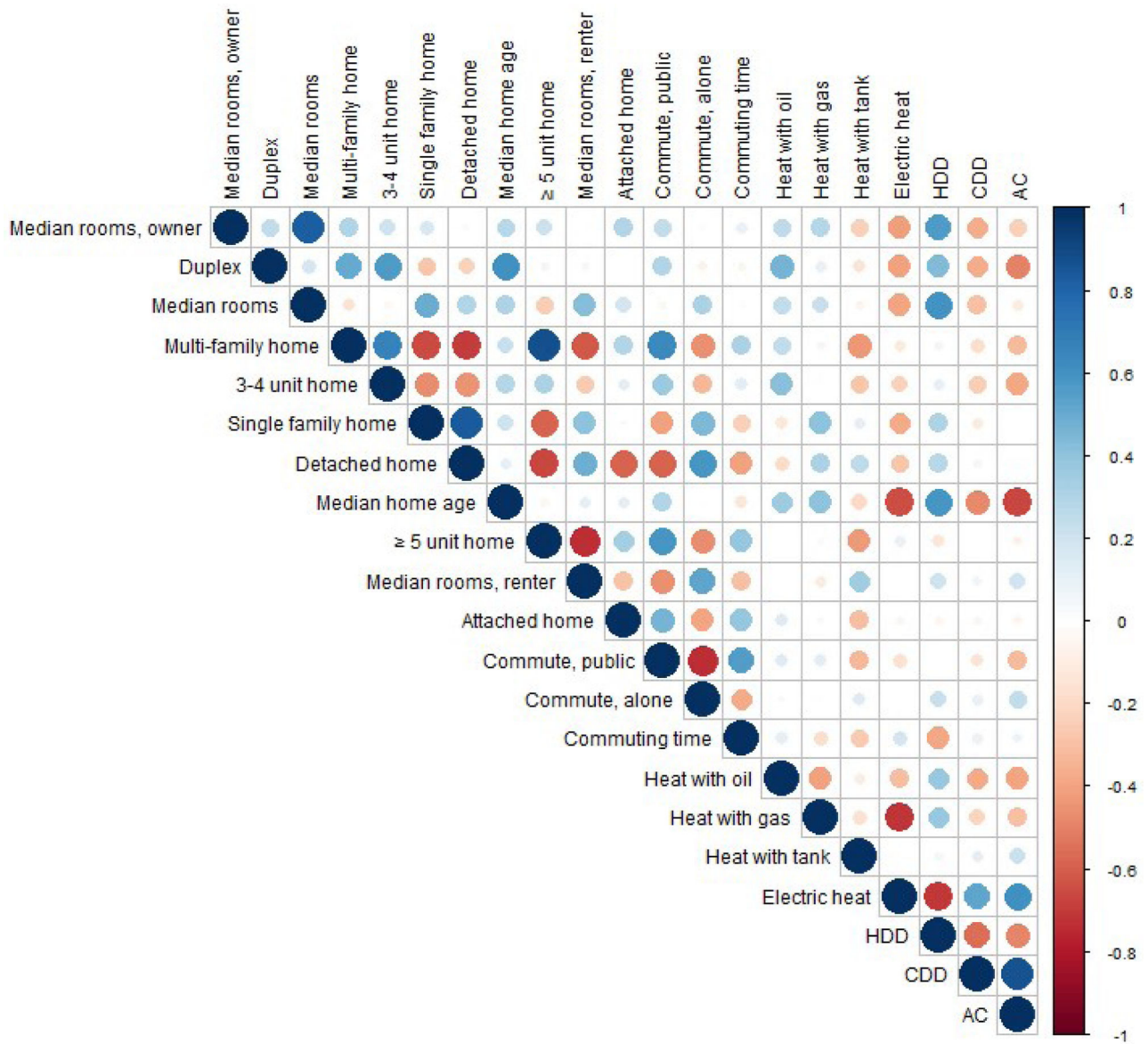
Figure 3:

Correlation of exposure factors[i] across CBSAs (n=312). Area of circles corresponds to absolute value of correlation coefficient, ordering of exposure factors in correlation matrix corresponds to ordering of variables in Table 4. "AC" variable refers to the full set of predicted and empirical values for prevalence of central A/C in the CBSA. See Supplemental Table 2 for numeric correlation coefficients.

[i]- abbreviation key: Median rooms, owner = Median number of rooms in residence, owner occupied home; Duplex = duplex homes; Median rooms = Median number of rooms in residence; Multi-family home = Multi-family homes (2 units in structure); 3-4 unit home = Homes with 3-4 units in structure; Single family home = Single family homes; Detached home = Detached single-family homes; Median home age = Median home age; 5 unit home = Homes with 5 units in structure; Median rooms, renter = Median number of rooms in residence, renter occupied; Attached home = Attached single-family homes; Commute,

public = Commuting on public transportation; Commute, alone = Commuting alone; Commuting time = Mean commute time in minutes; Heat with oil = Heating with oil; Heat with gas = Heating with utility gas; Heat with tank = Heating with tank gas; Electric heat = Heating with electricity; HDD = Heating degree days (annual sum of heating degree days (degrees below 65°F)); CDD = Cooling degree days (annual sum of heating degree days (degrees above 65°F)); AC = Prevalence of central air conditioning (predicted and empirical)

Table 1:

Exposure factors analyzed across CBSAs (n=312). All values used were average for the CBSA. The 2000 U.S. Census was the source of data for all variables, with the exception of the meteorological factors which were constructed using AQS data.

Exposure factors		Median (IQR) across CBSAs	CV across CBSAs
Housing characteristics	[*], [+]Median home age	28 years (11)	28
	[*]Detached single-family homes	65% (10)	13
	[*]Attached single-family homes	3% (3)	104
	[*]Single family homes	69% (8)	10
	[*]Duplex homes	3% (2)	66
	[*]Homes with 3-4 units in structure	4% (2)	42
	[*]Homes with 5 units in structure	12% (8)	48
	[*]Multi-family homes (2 units in structure)	21% (9)	37
	[*], [+]Median rooms in residence	5.4 (0.4)	6
	[*]Median rooms in residence, owner occupied	6.0 (0.5)	7
	[*]Median rooms in residence, renter occupied	4.2 (0.3)	6
Commuting	[*]Commuting alone	80% (5)	6
	[*]Commuting on public transportation	1% (1)	165
	[*]Mean commute time	22 min. (4)	15
Household heating	[*]Utility gas	52% (37)	48
	[*]Tank gas	6% (6)	73
	[*]Electricity	27% (35)	71
	[*]Oil	1% (7)	191
Meteorological factors	[1], [*], [+]Cooling degree days	3213 degree-days (1517)	31
	[2], [*]Heating degree days	6815 degree-days (3477)	33
	[+]Average annual relative humidity	69% (6)	11
Poverty	[+]Families below poverty level (of total families)	8% (5)	44

[1]Annual sum of cooling degree days (degrees above 65°F); n=311 for CDD

[2]Annual sum of heating degree days (degrees below 65°F)

[*]Analyzed as a predictor of heterogeneity in the association between $PM_{2.5}$ and mortality

[†]Included in machine learning analysis to predict prevalence of central air conditioning in CBSAs where data was missing

Table 2:

Central air conditioning prevalence across CBSAs. Prevalence used was the mean for each CBSA.

		Median (IQR) across CBSAs	CV across CBSAs	n
Observed data[1]	Central A/C	81% (39)	35	54
Predicted data[2]	Central A/C	72% (28)	23	257
All (Observed + Predicted data)	Central A/C	73% (29)	26	311[3]

[1] Data from national surveys

[2] Data predicted from machine learning approach

[3] Data could not be predicted for 1 CBSA due to a missing CDD value

Table 3:

Results of univariate regression of exposure factors as a predictor of heterogeneity in the association between a 10 μg/m^3 increase in PM$_{2.5}$ and percent change in mortality, per IQR change in exposure factor (n=311)[1]

Exposure factors		Beta (5th; 95th %tile)	F-statistic	p-value
Housing characteristics	Median number of rooms in residence, owner occupied	0.258 (0.120; 0.396)	13.41	<0.001
	Duplex homes	0.133 (0.061; 0.204)	13.29	<0.001
	Median number of rooms in residence	0.102 (0.018; 0.186)	5.616	0.018
	Multi-family homes (2 units in structure)	0.087 (0.009; 0.164)	4.837	0.029
	Homes with 3-4 units in structure	0.107 (0.011; 0.203)	4.773	0.030
	Single family homes	−0.081 (−0.165; 0.003)	3.605	0.059
	Detached single-family homes	−0.081 (−0.167; 0.006)	3.329	0.069
	Median home age	0.08538 (−0.046; 0.217)	1.625	0.203
	Homes with 5 units in structure	0.052 (−0.039; 0.143)	1.263	0.262
	Median number of rooms in residence, renter occupied	0.034 (−0.053; 0.120)	0.574	0.449
	Attached single-family homes	0.009 (−0.035; 0.053)	0.152	0.697
Commuting	Commuting on public transportation	0.017 (−0.002; 0.035)	3.172	0.076
	Commuting alone	−0.014 (−0.074; 0.045)	0.222	0.638
	Mean commute time	−0.010 (−0.127; 0.108)	0.026	0.872
Household heating	Oil	0.148 (0.094; 0.202)	28.93	<0.001
	Utility gas	−0.412 (−0.619; −0.205)	15.23	<0.001
	Tank gas	−0.107 (−0.391; 0.178)	0.540	0.463
	Electricity	0.073 (−0.143; 0.290)	0.444	0.506
Meteorological factors	[1]Heating degree days	0.319 (0.144; 0.494)	12.77	<0.001
	[2]Cooling degree days	−0.188 (−0.367; −0.010)	4.264	0.040
Complete A/C (predicted + empirical) (n=297)	Central A/C	−0.136 (−0.291; 0.018)	2.992	0.085

[1]Data could not be predicted for 1 CBSA due to a missing CDD value

[2]Annual sum of heating degree days (degrees below 65°F)

[3]Annual sum of cooling degree days (degrees above 65°F)

Table 4:

Results of multivariate regression of exposure factors as a predictor of heterogeneity in the association between a 10 μg/m^3 increase in PM$_{2.5}$ and percent change in mortality, per IQR change in exposure factor. Kitchen sink method was used to drop one covariate at a time based on highest p-value, until all covariates included were significant.

Exposure factors	Beta (5th; 95th %tile)	p-value
[1]Model 1		
Median home age	−0.288 (−0.516; −0.059)	0.014
Duplex homes	0.135 (0.021; 0.249)	0.021
Median number of rooms, renter occupied	−0.123 (−0.243; −0.002)	0.047
Utility gas	−0.471 (−0.684; −0.257)	<0.001
Heating degree days	0.448 (0.162; 0.733)	0.002
Cooling degree days	−0.228 (−0.432; −0.023)	0.030
[2]Model 2		
Duplex homes	0.128 (0.059; 0.198)	<0.001
Median number of rooms, renter occupied	0.215 (0.081; 0.348)	0.002
Utility gas	−0.427 (−0.627; −0.227)	<0.001

[1]Adjusted R-squared: 0.126; F-statistic: 8.47

[2]Adjusted R-squared: 0.1144; F-statistic: 14.35