U.S. Environmental Protection Agency

Office of Research and Development

National Health and Environmental Effects Research Laboratory

*Division*

*Branch*

Quality Assurance Project Plan (QAPP) Title:

Influence of Exposure Differences on City-to-City Variations in PM2.5: Mortality Effect Estimates

QA Track ID:

*[Put QA track id here]*

QA Category:

A  B

|  |  |
| --- | --- |
| ORD National Program: |  |
| Product ID: |  |
| Output ID: |  |

|  |  |
| --- | --- |
| Research Product Deliverable ID: |  |
|  | | |

Approvals

|  |  |  |
| --- | --- | --- |
| Prepared by: |  | NHEERL/Div/Br |
|  | Name Date | Organization |
| Supervisor: |  | NHEERL/Div/Br |
|  | Name Date | Organization |
| QA Manager: |  | NHEERL/Div/Br |
|  | Name Date | Organization |

Table of Contents

[Table of Contents 2](#_Toc14337430)

[Revision History 3](#_Toc14337431)

[Section A – Executive Summary 3](#_Toc14337432)

[A.1 Introduction and Background 3](#_Toc14337433)

[A.2 Research Approach Summary 3](#_Toc14337434)

[A.3 Project Team, Roles, and Responsibilities 4](#_Toc14337435)

[A.4 Timeline 4](#_Toc14337436)

[A.5 Assessment / Oversight 4](#_Toc14337437)

[Section B - Experimental Design 4](#_Toc14337438)

[B.1 Sample/Data Collection, Gathering, or Use 4](#_Toc14337439)

[B.1.1 Physical Sampling and Analysis 4](#_Toc14337440)

[B.1.2 Environmental Information, Qualitative Information, and Surveys 4](#_Toc14337441)

[B.1.3 Existing Data 4](#_Toc14337442)

[Exposure Factors 5](#_Toc14337443)

[Meta-regression 5](#_Toc14337444)

[Quality Metrics 5](#_Toc14337445)

[B.1.4 Software and Application Development (includes tools and databases) 6](#_Toc14337446)

[B.1.5 Model Development 6](#_Toc14337447)

[B.1.6 Model Application 6](#_Toc14337448)

[B.1.7 Design, Construction, and Operation of Environmental Technologies (e.g., pilot plants) 6](#_Toc14337449)

[B.2 Data Analysis / Statistical Design / Data Management 6](#_Toc14337450)

[Exposure Factors 7](#_Toc14337451)

[Prediction of Central Air Conditioning 8](#_Toc14337452)

[Meta-regression 8](#_Toc14337453)

[Data Management 8](#_Toc14337454)

# Revision History

## Table 1. QAPP revision history.

|  |  |  |
| --- | --- | --- |
| **Revision Number** | **Date Approved** | **Revision** |
| 0 | 01/5/2019 | New document |
| 1 |  |  |

# Section A – Executive Summary

A.1 Introduction and Background

The inability to explain the city-to-city heterogeneity, both nationally and within a region, in PM2.5 mortality risk estimates observed in multi-city studies remains a key uncertainty in the examination of the relationship between short-term PM2.5 exposures and mortality. One potential reason for these differences is the use of central-site monitors as a surrogate for exposure which may introduce bias into the observed risk estimates if the central-site monitor-exposure relationship varies by city.

A previous study (Neas et al. 2017) determined the PM2.5 mortality risk estimates for 312 Core Based Statistical Areas (CBSAs) from 1999-2005. The objective of this analysis is to evaluate potential exposure factors as determinants of the heterogeneity in CBSA-specific associations between PM2.5 and mortality. These factors are related to housing characteristics, household heating, commuting to work, and meteorological factors. Another potentially important factor is the prevalence of air conditioning. Of the 312 CBSAs included in the analysis, only 54 of these included data on air conditioning prevalence. Machine learning techniques will be used to predict prevalence of central AC for the 258 CBSAs where AC data was not available. Meta-regression will be utilized to determine whether these CBSA-varying exposure factors result in differential mortality associations with PM2.5.

This work is part of task PEP 1.4: Health Effects of Multiple Air Pollutants across Multiple US Communities: Determinants of Local and Regional Heterogeneity under the Air, Climate, and Energy Program. Much of previous work and underlying data sources supporting this work is described in detail in QAPP-NHEERL-H/EPHD/EB/LMN/07-001-007 and QAPP-NHEERL/RCPS/LKB/2016-01-r0. The machine learning technique is covered in QAPP (D-CED-0030209).

A.2 Research Approach Summary

The generation of the health associations and the data used is covered under (QAPP-NHEERL-H/EPHD/EB/LMN/07-001-007) and in Neas et al. 2017. Briefly, the association between daily PM2.5 concentrations and non-accidental mortality in metropolitan areas was determined for 312 multi-county metropolitan areas across the continental United States for the years 1999-2005 using Poisson time-series models. The definitions for the 312 multi-county metropolitan areas in this analysis are based on the CBSA of the White House Office of Management and Budget. Health events were sorted into multi-county metropolitan regions centered on existing air quality monitors according to either county of event or county of residence. Meta-regression will be applied to the CBSA-specific risk estimates as described in QAPP-NHEERL/RCPS/LKB/2016-01-r0 to determine whether human exposure factors can explain the observed heterogeneity in the risk estimates.

A.3 Project Team, Roles, and Responsibilities

Dr. Laura Fields of the National Health and Environmental Effects Laboratory will serve as the primary investigator of this analysis.

Dr. Kimberly Kitsinis of the National Exposure Research Laboratory is responsible for the statistical analysis and machine learning.

Dr. Vanessa Shields of the National Center for Computational Toxicology is a post-doctoral fellow providing guidance on the machine learning techniques.

Dr. Luke Perry of the National Health and Environmental Effects Laboratory is the senior investigator of the Multi-city/Multi-pollutant Study.

A.4 Timeline

This study is expected to run from August 2019 to March of 2020 and will result in a peer-reviewed journal articles in FY20.

## A.5 Assessment / Oversight

Publications resulting from this project will undergo ORD clearance in STICS and peer-review of the journal article(s). For QA Category B projects, no additional assessments are required to be conducted by ORD QA Managers, although it may be included in Laboratory Competency Audits or conducted as directed by management.

# Section B - Experimental Design

## B.1 Sample/Data Collection, Gathering, or Use

As detailed in QAPP-NHEERL-H/EPHD/EB/LMN/07-001-007 and Neas et al, health events were aggregated into time-series of Poisson counts by date of event within each category and metropolitan region, matched with air quality and meteorological data by date and region. This was then analyzed using a generalized additive model to estimate the association between time-varying health event counts and time-varying air quality measure(s) while adjusting for time-varying confounders such as trend, season, and meteorology.

Air quality measurements for multiple U.S. cities have been obtained from EPA’s Air Quality System (http://www.epa.gov/airdata/) Data Mart, including data from the State and Local Air Monitoring System, the National Core Monitoring System, the Chemical Speciation Network and the IMPROVE Network. The data include information on the composition of size-specific particulate matter (PM) mass/components and on gaseous co-pollutants, including ozone, nitrogen dioxide, sulphur dioxide, and carbon monoxide. The focus of this project was on the PM¬2.5 concentrations.

The detailed, individual-level mortality data for the entire U.S. from 1985 through 2005 have been obtained through the National Center for Health Statistics (NCHS) from administrative systems for vital event records maintained by State and local health departments (http://www.cdc.gov/nchs/about.htm). This analysis only included mortality data from 2001-2005.

Daily meteorologic data for multiple U.S. cities has been obtained from the U.S. Department of Commerce’s National Climatic Data Center (http://www.ncdc.noaa.gov/oa/ncdc.html). The daily data include 24-hour averages of ambient temperature, dewpoint temperature, apparent temperature, relative humidity, cooling degree days, heating degree days, and barometric pressure; and total 24-hour precipitation.

#### Exposure Factors

A number of exposure factors related to housing characteristics, household heating, commuting to work, and meteorological factors were constructed and analyzed as potential predictors of heterogeneity in the association between PM2.5 and mortality. As described in QAPP-NHEERL-H/EPHD/EB/LMN/07-001-007 this data was acquired from the American Housing Survey (AHS), available from the Department of Housing and Urban Development's website (<http://www.census.gov/programs-surveys/ahs/data.html>).

In addition to the exposure factors described above, a metric/surrogate of air exchange rate was desired to be included as a potential predictor of heterogeneity in the association between PM2.5 and mortality. Of the 312 CBSAs included in the analysis, 54 of these included data on air conditioning prevalence within the CBSA from the AHS. Machine learning techniques were used to predict prevalence of central AC for the 258 CBSAs where AC data was not available. Figure 1 in QAPP (D-CED-0030209) presents a workflow for deciding a machine learning model (see below).

Following the workflow we have more than 50 samples and want to predict a quantity rather than a category leading us to the models in the regression group (upper right circle). We have less than 100K samples. We are not sure if few features are important so we have chosen a modeling approach from each branch. Based on this work flow, the machine learning techniques chosen for this analysis include lasso and linear support vector machines (SVR (kernel=linear)). This workflow is not inclusive of all of the techniques available. We will also test random forest models. This approach tends to be one of the most common machine learning techniques that is relatively easy to implement and interpret.

#### Meta-regression

Meta-regression will be applied to the CBSA-specific risk estimates as described in QAPP-NHEERL/RCPS/LKB/2016-01-r0 to determine whether human exposure factors can explain the observed heterogeneity in the risk estimates.

#### Quality Metrics

As detailed in the air quality, health, meteorological data, housing data have been subjected to quality reviews by organizations outside of ORD.

The US Environmental Protection Agency (EPA) is the Federal Agency with statutory responsibility for the collection and processing of air quality data. Air quality measurements for multiple U.S. cities have been obtained from EPA’s Air Quality System (<http://www.epa.gov/airdata/>) Data Mart. The air quality data was reviewed by the various state and local air quality monitoring agencies and by EPA’s Office of Air Quality, Planning and Standards. Quality assurance documentation for the Air Quality System may be found at <http://www.epa.gov/ttn/amtic/qalist.html>.

The National Center for Health Statistics (NCHS) within the Department of Health and Human Services is the Federal Agency with statutory authority to the collection and processing of mortality data. The detailed, individual-level mortality data for the entire U.S. from 1985 through 2005 have been obtained through NCHS from administrative systems for vital event records maintained by State and local health departments (http://www.cdc.gov/nchs/about.htm).

Metadata for the 21 annual mortality data files has already been prepared by NCHS and copies of this metadata reside at \\AA\ORD\RTP\NERL\_Multicity\METADATA\NCHS Documentation Files\. The mortality data was reviewed by the various state and local vital event registries and the NCHS, including a recoding of nosology codes for a sample of the death records. Quality assurance documentation of NCHS operating procedures may be found at <http://www.cdc.gov/nchs/nvss/mortality_methods.htm>.

The National Climatic Data Center within the U.S. Department of Commerce is the Federal Agency with statutory responsibility for the collection and processing of meteorologic data. Daily meteorologic data for multiple U.S. cities has been obtained from the NCDC (<http://www.ncdc.noaa.gov/oa/ncdc.html>), including 24-hour averages of ambient temperature, dew point temperature, apparent temperature, relative humidity, cooling degree days, heating degree days, and barometric pressure; and total 24-hour precipitation. Metadata for the original data files has already been prepared by NCDC and resides at https://www.ncdc.noaa.gov/data-access. The meteorological data was reviewed by the National Weather Service and by the NCDC. Quality assurance documentation of NCDC operating procedures may be found at <http://www.ncdc.noaa.gov/oa/climate/ghcn-daily/index.php?name=quality>.

The Bureau of the Census within the U.S. Department of Commerce is the Federal Agency with statutory responsibility for the collection and processing of population and housing data. Population and housing data has been obtained from the Census (<http://www.census.gov/data.html>), including population/housing counts and characteristics for censual years (1990, 2000, and 2010) and population estimates for intercensal years. Acquisition of Census Data will follow the the SOP (\\AA\ORD\RTP\NERL\_Multicity\METADATA\). The data from the American Housing Survey have been reviewed y the U.S. Bureau of Census. Quality assurance documentation of NCDC operating procedures may be found at <http://www.census.gov/about/policies/quality/standards.html>.

### B.1.4 Software and Application Development (includes tools and databases)

### B.1.5 Model Development

### B.1.6 Model Application

### B.1.7 Design, Construction, and Operation of Environmental Technologies (e.g., pilot plants)

## B.2 Data Analysis / Statistical Design / Data Management

The health effect estimates used in this analysis were generated as part of a previous publication (Neas et al. 2017)

#### Exposure Factors

The table below lists the exposure factors that will used for this analysis. The factors were based on their previous studies relationships with personal exposures (QAPP-NHEERL/RCPS/LKB/2016-01-r0).

|  |  |  |  |
| --- | --- | --- | --- |
| **Housing characteristics** | **Commuting** | **Household heating** | **Meteorological factors** |
| Median home age (years) | Mean commute time (minutes) | Utility gas (%) | Annual sum cooling degree days |
| Detached single-family homes (%) | Commuting alone (%) | Tank gas (%) | Annual sum of heating degree days |
| Attached single-family homes (%) | Commuting on public transportation (%) | Electricity (%) |  |
| Single family homes (%) |  | Oil (%) |  |
| Duplex homes (%) |  | Other (wood, coal, solar, other fuel, or no heating fuel; %) |  |
| Homes with 3-4 units in structure (%) |  | Utility gas (%) |  |
| Homes with ≥ 5 units in structure (%) |  | Tank gas (%) |  |
| Multi-family homes (≥ 2 units in structure; %) |  |  |  |
| Median number of rooms in residence |  |  |  |
| Median number of rooms in residence, owner occupied |  |  |  |

Based on the AHS data a number of exposure factors will be constructed. The median home age will be calculated as: median home age = 2000 – median year the home was built, to approximate the home’s age in the middle of the study span (1999-2005). To calculate the percentage of each housing type within a CBSA (i.e., attached single family homes, detached single family homes, duplex homes, and homes with 3-4 units in the structure), the corresponding variable for number of housing units of each type will be divided by the total number of housing units in the CBSA. Larger categories of housing type (i.e., single family homes (overall), homes with ≥ 5 units in the structure, and multi-family homes) will be calculated by summing the appropriate individual variables, and then dividing the total by the number of housing units in the CBSA. The percentage of individuals in the CBSA commuting alone, and the percentage commuting on public transportation (including taxi cabs) will be calculated by dividing the number of individuals commuting by that mode by the total base population of individuals with mode of commute data. Percentage of homes using a specific type of heating fuel (i.e., utility gas, tank gas, electricity, or oil) in each CBSA will be calculated by dividing the number of homes using that type of heating fuel by the base population for the heating fuel variables. To calculate percentage of homes using a different type of heating fuel (‘other’), the total sum of homes using wood, coal, or solar fuel for heating, using a heating fuel which was not indicated, or using no heating fuel, will be divided by the base population for the heating fuel variables. Heating degree days (HDD) and cooling degree days (CDD) are variables from the NCDC and were summed over the study year period for each CBSA. All other variables were taken directly from the original data source without transformation.

#### Prediction of Central Air Conditioning

Exposure factors known to be predictive of central AC prevalence will be chosen a priori. These include housing characteristics and household heating derived from the AHS, meteorological variables (annual average temperature, annual average dewpoint temperature and annual average relative humidity) from NCDC, and a poverty related variable (prevalence of families below poverty level variables. The construction of the poverty related variables are described in more detail in SOP *Heterogeneity Analysis Methods 2015 11 23.pdf.* Briefly, the prevalence of families below the poverty level is equal to families below the poverty level divided by the total number of households.

These factors will be used to train and test lasso, SVM, and random forest models to predict prevalence of central AC by CBSA. Model validation is also important. 10-fold internal and 4-fold external cross-validations techniques will be employed. As further detailed in QAPP (D-CED-0030209), mean root mean square errors (RMSE) and mean R2 will be examined to determine the best model. If all models are similar a consensus model can be used that average across the results of all models. Machine learning models will be implemented in R (version 3.1.2) using the package ‘caret’ available on CRAN.

#### Meta-regression

Meta-regression will be employed to assess the determinants of this heterogeneity including ecological-level characteristics of the CBSAs. The format of the dependent variable will be the beta divided by the standard error. In other words, the inverse-weighted variance of the log rate ratios. This is to ensure less importance is placed on those estimates that are the most uncertain. We will use the p-values to determine whether the exposure factor significantly explains the variation in health effect estimates.

#### Data Management

As described in QAPP-NHEERL-H/EPHD/EB/LMN/07-001-007, environmental data (air quality and meteorology) have been downloaded to and retained on a secure EPA network drive (L:\Lab\NERL\_Multicity) administered by Luke Perry (NHEERL). All the machine learning codes and data sets used will also be stored on this drive. Access to this drive requires both an EPA user name with a valid password and prior permission from the drive administrator.

Health data (mortality and hospitalization) at the individual level along with documentation are retained on a secure EPA network drive (L:\PRIV\US\_Deaths) administered by Luke Perry (NHEERL). The original physical media (CDs and external hard drives) used to transfer the data to EPA are retained in Room 50 of the EPA Human Studies Facility (HSF). Backup copies of working data files, computer programs, and key results will also be retained on this drive.

Computer program development and execution of the health models along with working copies of necessary data files will be retained on a T7600 computer in the EPHD GIS Laboratory (Room 111a of the EPA HSF). Access to the computer requires an EPA user name with a valid password. Computations of the meta-regression will be conducted and preserved on the secure EPA network drive.