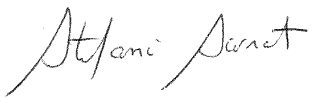
Quality Assurance Project Plan for Project 4

EPA Clean Air Research Center

Southeastern Center for Air Pollution and Epidemiology

Emory University and Georgia Institute of Technology

Management Approvals:



10/6/11

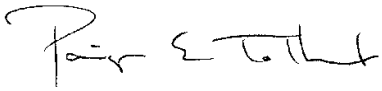
Stefanie Sarnat — Date
Project 4 Principal Investigator

10/10/11

Mitch Klein — Date
Quality Assurance Advisor

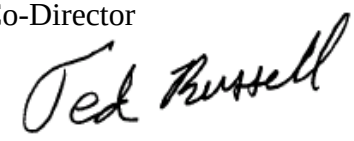
Guangxuan Zhu — Date
Quality Assurance Advisor

Kate Hodgins — Date
Quality Assurance Manager



10/6/11

Paige Tolbert — Date
Co-Director



10/5/11

Ted Russell — Date
Co-Director

Quality Assurance Project Plan for Project 4

EPA Clean Air Research Center

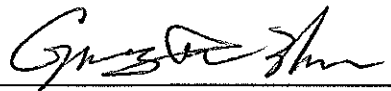
Southeastern Center for Air Pollution and Epidemiology

Emory University and Georgia Institute of Technology

Management Approvals:

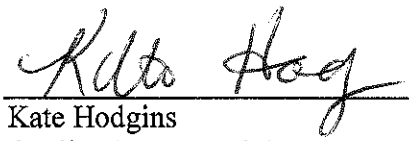
Stefanie Sarnat — Date
Project 4 Principal Investigator

Mitch Klein — Date
Quality Assurance Advisor



10/8-2011

Guangxuan Zhu — Date
Quality Assurance Advisor



10/8/11

Kate Hodgins — Date
Quality Assurance Manager

Paige Tolbert — Date
Co-Director

Ted Russell — Date
Co-Director

Quality Assurance Project Plan
For Project 4

EPA Clean Air Research Center

Southeastern Center for Air Pollution and Epidemiology

Emory University and Georgia Institute of Technology

1. Background/Purpose

Substantial epidemiologic evidence exists to support an association between ambient air pollution and acute cardiorespiratory health effects.[1-4] Numerous time-series studies have shown positive associations between the major ambient air pollutants [including ozone (O_3), carbon monoxide (CO), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), PM_{10} and $PM_{2.5}$ (particulate matter with aerodynamic diameter less than 10 and 2.5 microns, respectively)] and cardiorespiratory conditions using mortality, hospital admissions, and ED visit data. However, several major research questions remain about the degree to which the findings of these studies are generalizable between locations and whether the observed health effects are due to the individual pollutants measured or to pollutants acting in combination with other pollutants.

The purpose of the Southeastern Center for Air Pollution and Epidemiology (SCAPE) – Project 4, A Multi-City Time-Series Study of Pollutant Mixtures and Acute Morbidity, is to clarify the impacts of air quality on acute cardiorespiratory severe morbidity in at least five US cities (including Atlanta, GA; St. Louis, MO-IL; Dallas, TX; Birmingham, AL; and 1-3 other cities, to be determined) using novel pollutant mixture characterization (MC) metrics. In particular, this study aims to explain apparent between-city heterogeneity in short-term associations between air quality measures and cardiorespiratory emergency department (ED) visits and hospital admissions (HA). This analysis will enable a comprehensive assessment of the generalizability of findings regarding the health impacts of pollutants and pollutant mixtures.

Specific Project Objectives

Objective 1. Characterize ambient air pollution mixtures and populations in multiple US cities.

A) Characterize the multi-pollutant atmospheres using complementary approaches:
   - Population-weighted average and spatially-resolved daily monitoring data
   - Single-species source tracers
   - $PM_{2.5}$ source appointment using chemical mass balance (CMB) approaches
   - Pollutant groupings using positive matrix factorization (PMF) and UNMIX
   - Modeled reactive oxygen species (ROS)
   - Community Multiscale Air Quality (CMAQ) pollutant and source outputs

B) Characterize the study populations using data on sociodemographics, housing characteristics, and morbidity.

Objective 2. Compare and summarize short-term associations between MC metrics and cardiorespiratory outcomes (ED visits and HA) across the cities, including a comprehensive assessment of lag effects.

Objective 3. Examine and explain heterogeneity of associations across cities by:

A) Applying multi-pollutant epidemiologic air pollution mixture models
B) Examining season- and temperature-specific effects
C) Assessing and comparing concentration-response functions
D) Assessing and comparing impacts of exposure measurement error
E) Assessing and comparing associations in susceptible and vulnerable populations

Methods for Data Acquisition to address Project Objectives

To address the three project objectives, we will primarily utilize historic secondary data on ED visits and HAs, US Census indicators, and air quality. Data will be obtained from publically available online data sources (e.g., the USEPA Air Quality System, Census websites) and from sources that agree to make data available for use in this project (e.g., the SouthEastern Aerosol Research and Characterization [SEARCH] network, individual participating hospitals or hospital associations). Some datasets will be acquired from in-house data warehouses at Emory and Georgia Tech, compiled under previous and ongoing studies, or will be acquired specifically for use in the current project. All data to be used in this project will be obtained in electronic format and no direct data entry will be required. Data will either be downloaded from online repositories or will be transferred to the study team via e-mail attachments, CD-ROMs, or ftp or Sharepoint sites. All data will be saved on the SCAPE password-protected network drive.

2. Design

Mixture Characterization Metrics

We will acquire rich air quality monitoring databases for each city, and utilize a number of modeling tools to develop complementary mixture characterization metrics for use in descriptive and epidemiologic analyses: (1) population-weighted averages and spatially-resolved concentrations; (2) single-species source tracer data; (3) $PM_{2.5}$ source apportionment outputs using CMB approaches; (4) factor analysis outputs using PMF and UNMIX; and (5) modeled ROS levels. By employing the same modeling approaches, inter-city comparisons of results that highlight similarities and differences in air quality will be facilitated. The SCAPE Air Quality Core will provide all air quality monitoring data and mixture characterization metrics to Project 4 for use in addressing the specific project objectives.

Descriptive Analysis

In collaboration with the SCAPE Air Quality Core personnel, we will conduct a thorough descriptive analysis of pollutant atmospheres in each city using the measured and modeled MC metrics. A similar descriptive analysis of population (sociodemographic factors) and housing characteristics (air conditioning prevalence, air exchange rates [AERs]) will be conducted to better understand potential sources of heterogeneity of city-specific health associations.

Main Epidemiologic Analysis

We will investigate and summarize short-term associations between MC metrics and cardiorespiratory outcomes in the five cities using a two-stage approach. We conceptualize the first stage as city-specific case-crossover analyses,[7] with referent periods selected by matching on same-day temperature within calendar month and year. In practice, we will analyze these data using Poisson generalized linear models (GLM), since the conditional logistic regression estimating equation that is used in time-stratified case-crossover analyses is equivalent to a Poisson time-series model with indicator variables for each stratum when the disease is rare and there is a shared exposure (such as air pollution).[8,9] Poisson models have an advantage over the conditional logistic regression models in that the variance of the parameter estimates can be scaled to account for overdispersion.[8]

In the second stage of analysis, we will estimate a summary effect and the heterogeneity of effects between cities using meta-regression and generalized least squares, as conducted by others.[5, 6] This two-stage analytic approach is useful as it allows for conducting case-crossover analyses and for obtaining city-specific effect estimates. Moreover, it can easily incorporate approaches applied in the first stage of analysis to address Objective 3a, such as methods for accommodating multicollinearity in multi-pollutant models (e.g., LASSO) and non-parametric approaches [e.g., Classification and Regression Tree (CART), random forests].

Criteria for Success

For the research to be successful, we must accomplish the following:

1. Obtain individual-level ED visit and HA data for each of the cities
2. Process these data into comparable analytical ED visit and HA databases for each city
3. Acquire ZIP code level US Census data of interest for each city
4. Develop the MC metrics of interest for each city (an Air Quality Core activity)
5. Link the health, Census, and air quality data by date and/or ZIP code
6. Conduct epidemiological analyses using appropriate statistical models
7. Disseminate results to the scientific community through presentations at scientific conferences and publication in peer-reviewed journals.

3. Data Gathering Methods

Health Outcome Data

Data Sources

Individual-level ED visit and HA data for each city will be obtained from electronic billing records requested from state hospital associations, state health departments, other state organizations, and/or individual hospitals. To address the project objectives successfully, the data in each city will be for time periods (ranging from 6.5 to 17 years for each outcome type across cities) that overlap with the detailed air quality data in each city and will include all visits to acute care hospitals located in counties that represent the metropolitan areas of each city made by patients residing in ZIP codes located wholly or partially in these same counties.

Logistical considerations will largely determine the specific data sources from which we will obtain data, given that generally only one organization per city collects and provides hospital billing data for research purposes. In the event that our data of interest are not available through the respective state hospital association, state health department, other organization, we will target individual hospitals directly to provide the required electronic billing data.

All health data acquired will be subject to extramural agreements and/or data use agreements with the participating organizations providing the data. These agreements will be routed to Emory's Office of Sponsored Programs and Office of Grants and Contract before the data are purchased and received by Emory.

Data Elements

The same data elements will be obtained from each data source in each of the study cities, and will include (if available) for each visit: a unique patient identifier (e.g., medical record number, or blinded social security number), unique visit number, admission and discharge date and time, admission source, admission type, primary and secondary International Classification of Diseases 9th Revision (ICD-9) diagnosis codes, Current Procedural Terminology (CPT) codes, age, date of birth, gender, race/ethnicity, method of payment (e.g., Medicare, Medicaid, self-pay), and address and/or ZIP code of patient residence. For ED visits, information on whether the visit resulted in an inpatient admission will also be obtained, if available.

Census Data

We will acquire ZIP Code Tabulation Area (ZCTA)-level Census data for describing and comparing sociodemographic and home characteristics among the five city populations. Variables of specific interest include age, race, high school graduation, median income, percent of the population in poverty and unemployed, and housing unit median number of rooms, value, and year built. We will also obtain American Housing Survey data to estimate air conditioning prevalence and for calculations estimating ZIP code specific air exchange rates in each city.

Air Quality Data (through the Air Quality Core)

This project will capitalize on two or more years of daily speciated air quality measurements in each city, provided by various local speciation monitors: USEPA Supersites in Atlanta, Pittsburgh, and St. Louis; the SouthEastern Aerosol Research and Characterization (SEARCH) network in Atlanta and Birmingham; Aerosol Research and Inhalation Epidemiological Study (ARIES) measurements in Atlanta, Dallas, Pittsburgh, and St. Louis; and the Assessment of the Spatial Aerosol Composition in Atlanta (ASACA) network in Atlanta. Data from USEPA STN monitors will supplement these local monitoring efforts in all cities for all study years. Routine monitoring data will also be obtained from the USEPA Air Quality System (AQS) and other local stations, and meteorological data will be obtained from the National Climatic Data Center (NCDC) and monitors associated with air quality sites. Specifically, hourly values of the gaseous pollutants O_3, CO, NO_2, NO_x, and SO_2 will be collected, as well as daily and hourly measures of PM_{10} and $PM_{2.5}$. Hourly and 24-hour PM composition data, including major ions (SO_4^{-2}, NO_3^-, NH_4^+), carbon fractions (EC, OC) and trace metals, are available from 3 to 10 monitors in each city (daily data from 1 to 4 sites/city). From the Supersites and ARIES measurements, additional detailed data are available, including daily speciated coarse PM, ultrafine PM, water-soluble metals, speciated particle-phase organics, and speciated VOCs in select cities. These air quality data have largely been assembled, or are being assembled, under our existing studies. These data will be utilized by the SCAPE Air Quality Core to create the specific mixture characterization metrics of interest to the current project. Please refer to the SCAPE QAPP for the Air Quality Core for details regarding quality assurance and quality control procedures to be followed for these data.

4. Data Quality

Data quality procedures for the health outcome data, Census data, and epidemiologic modeling are detailed in this section. Many of these data are already used in our ongoing projects and undergo periodic, extensive validation and auditing. There are no constraints on the existing data that affect its use in the proposed project.

Health Outcome Data

The health outcome data used in this study will consist of historic records of ED visits and HAs. We will apply rigorous qualitative and quantitative measures of quality assessment for these data.

For the qualitative assessment of the ED and HA data, we will rely on the judgment of the individual hospital data managers regarding its limitations based on data entry procedures, storage restrictions in the source system, accuracy and reliability of the information in each data field, and any peculiarities in the data resulting from the method of data extraction from the source system. Upon receipt of the data at Emory, we will also apply qualitative assessments by following standard data check procedures on each dataset, and recording all pertinent information on the "*SCAPE Project 4 Health Data Check Form.docx*" (attached). Datasets will be imported to SAS Version 9.2 (SAS Institute Inc., Cary, NC) for this process, and initial data cleaning will be conducted to facilitate data characterization (including formatting of ICD-9 codes and assigning hospital names and identifiers)

Following the initial SAS import and data check, data sets from each source will be run through a standard cleaning code, customized as needed. All data cleaning procedures will be guided by and documented via the "*SCAPE Project 4 Health Data Preparation Procedures Form.docx*" (attached). Briefly, the standard cleaning code will include steps to read in and select data of interest; clean patient identifiers, demographic variables, ICD-9 diagnosis and procedure codes, and ZIP codes; calculate length of stay (for HA data, if applicable); create ZIP code groups (e.g., that delineate the study area and any spatial subsets of interest), delete invalid observations (in the second round of cleaning, as described below), create within patient repeat visit counter variables, fully label the dataset, create case (e.g., respiratory and cardiovascular disease groupings) and control groups of interest, and finally aggregate valid data to daily counts of specific outcomes and population subsets of interest. Random checking of data in final datasets will be conducted, by comparing to raw data, to ensure that these datasets were created properly. Potential data errors and invalidity will be discussed among members of the entire research team. Applying a standard cleaning program across all datasets will help to ensure consistency and comparability in health outcome data across cities to the extent possible given the acquired data.

During and after initial data checks and cleaning procedures are completed, quantitative assessments will be conducted to validate the data. Quantitative assessment of the ED and HA data will include examining the raw data using descriptive statistics. The daily counts for all visits and case groups of interest will be evaluated for unusual day-to-day variability using frequency tables and univariate statistics. Any within hospital trends that appear inconsistent (e.g., days with counts greater than or less than 3*standard deviation, or time periods for which

the daily counts appear inconsistent compared to the remainder of the time) will be invalidated, particularly if no reasons (e.g., merging of two hospitals, or change in coding practices) for such trends can be identified. Correlations between daily counts of case groups and all ED visits or HAs among the different hospitals will also be examined. Frequencies of case groups will be assessed to evaluate consistency between similar types of hospitals. All invalidations will be discussed with the QA Advisor and Project 4 QC Reviewer, and all discussions and information pertaining to each invalidation will be documented. Through an iterative procedure, the standard cleaning program will be re-run to take into account any data invalidations deemed appropriate. The list of invalidations and an overall hospital entry/exit listing will be utilized in the epidemiologic modeling component of the project to help account for time trends in the outcome data.

Throughout the data check and data cleaning process, the following QC parameters will be characterized for the final ED visit and HA datasets for each city:

- The *representativeness* of the temporal and spatial extent of study domain, including # of hospitals included in the dataset and the estimated % of visits captured by the acquired dataset (pre- and post- data validation).
- The *data completeness* will be recorded by hospital and by city, with information on missingness for each data element.
- The *comparability* of visit counts to published and/or publically available aggregate data will be determined. For example, the annual number of ED visits will be compared to reports of annual visits for each hospital published by respective State Health Departments.

Because these are secondary data, with no feasible opportunity for re-acquiring the data, we will not state limits that will constitute failure among these parameters, however the information from these QC parameters will provide useful information for final epidemiological results interpretation. For example, our city-specific health datasets may exclude visits from certain hospitals, which may impact the *representativeness* and *data completeness* for the city of interest, as well as the *comparability* of our observed visit counts to those published by others. Data may be missing from specific hospitals for several reasons, including the hospital's lack of relationship with our primary data source, lack of willingness to participate in our study if approached directly, or inability to provide us with the specific data that we require. The consequence of such missing data may result in our inability to fully describe and characterize healthcare utilization of the entire population of the city, with impacts on both spatial representativeness as well as subpopulation representativeness (e.g., if an excluded hospital serves a specific portion of the population, such as children). Differential missingness across cities may be a reason for observing heterogeneity in air pollution health risk estimates, and will be important in guiding specific analyses and final interpretation of our epidemiologic results.

US Census Data

ZCTA-level Census data will be obtained from publically-available online data repositories for each county of interest to the project. We will defer to US Census Bureau descriptions of Census data limitations to accuracy and representativeness. To ensure that the data we have acquired are as accurate as they can be, for a subset of ZIP codes and counties, we will download data from a

second website and make sure that the numbers obtained are comparable. In the event that there is more than a 10% discrepancy in the numbers obtained from each site, a further examination will be conducted to better understand the source of discrepancy and to re-download the data as needed.

Data Quality Assessments

The QC Reviewer will be responsible for documenting and keeping records of QA/QC procedures and any deficiencies that occur. As part of routine Data Quality Assessments, the QC Reviewer will ensure that the *SCAPE Project 4 Health Data Check Form* and *SCAPE Project 4 Health Data Preparation Procedures Form* are followed for each health dataset utilized in this project, that data invalidations are documented and justified, and that the QC parameters (*representativeness*, *data completeness*, and *comparability*) are computed and characterized for each dataset and corresponding city. The QC Reviewer will send all completed forms and information on QC parameters to the Center's QA Manager.

5. Data Reduction

As described above, the health outcome data for each city will be transformed into a standard configuration and aggregated to daily counts of visits, for both ED visit and HA data separately. Each health data series will be validated according to the procedures described above and non-valid data will be removed. The data reduction process will be documented according to the "*SCAPE Project 4 Health Data Preparation Procedures Form.docx*". The outcome data will then be combined with daily (and, in some cases, ZIP code specific) air quality data and census data to form the final analytic files. Descriptive statistics and multiple regression models will be utilized to address the project objectives. SAS Version 9.2 (SAS Institute Inc., Cary, NC) and/or R software will be used for all statistical analyses. Microsoft Excel 2007 and ArcGIS 9 will be used for graphical display of the results. Results will be presented as summary statistics (e.g., means, standard deviations, ranges) as well as relative risks (and 95% confidence intervals) from epidemiological models. The assessment of data quality, discussed above, will be utilized to aid in the interpretation of these results.

6. Interaction of the Players

The PI of the project is Stefanie Sarnat. She will direct the project and has the authority to change how the various players interact. She is supported by Co-PIs from Emory (Paige Tolbert) and Georgia Tech (James Mulholland), who will facilitate interaction between this project and the Cores. Lyndsey Darrow, Andrea Winquist, and Mitch Klein are co-investigators on the project; doctoral and masters level students (to be determined) will also be involved. Priya Kewada, the project coordinator will serve as the QC Reviewer once data collection activities start. The QC Reviewer will be responsible for documenting and keeping records of QA/QC procedures and any deficiencies that occur and for reporting routinely to the QA Manager. The efforts of the project will also be overseen by the Core's QA Advisor, Guangxuan Zhu, who will work with the PI and Project Coordinator to conduct annual reviews of project methods and QA

procedures. All investigators and study personnel with access to health data will maintain their CITI certification by taking refresher courses every-other-year (or as specified by Emory University). The PI will work closely with the QC Reviewer and all study personnel involved in data processing and analysis to ensure data quality procedures are followed appropriately.

This project builds on the investigators' ongoing research efforts to examine the association between air pollution and morbidity in Atlanta. Much of this work has been conducted within the framework of the ARIES study and with the consultation of the study's multi-disciplinary advisory committee. We will solicit feedback from this committee as well as the SCAPE Scientific Advisory Committee (SAC) to ensure that the methods and designs used in this study optimally address the proposed hypotheses and are consist with the approaches utilized in our previous research efforts. Likewise, study results will be presented to the SAC, providing a valuable opportunity to receive external evaluation prior to manuscript submission. A formal peer review of the study, prior to data publications, will include review of all programs for data quality checking, data management, and data analysis. Emory University will assign applicable in-house expertise to this activity with available project funds. This process will ensure that all data analysis summaries (e.g. descriptive statistics, effect estimates, and 95% confidence intervals, etc) from which inference and conclusions are made, as well as any resulting publications are error-free. Final evaluation will also be achieved in assessing whether the data collected successfully addressed the specific hypotheses and through the submission of manuscripts or peer reviewed scientific journals.

7. Data Management

Health data will be transferred to Emory using various methods, including CD-ROM, zip disks, and/or e-mail attachments. The data will immediately be placed on our password-protected secure network drives that are regularly backed up, and the disks (if any) will be stored in a locked filing cabinet located in a locked office. Access will be limited to those who are on the Emory IRB for the project and have approval to work on the project as determined by the PI. The air quality data will be hosted on the Georgia Tech website and transferred as needed to the Emory network for use on this project.

An official listing of all secondary data sources utilized in this project will be kept by the QC Reviewer, and will be referred in the event of presentation or publication of results, so that appropriate identification of the data sources may be made

Final health datasets will be stored as "read only" on the network drive; modifications will be made to copies of these datasets as needed. The tasks of the QC Reviewer that relate to data management include:

1) Check to ensure that our Emory network drive has the most current versions of all the different air quality datasets that are on the Georgia Tech website.
2) Check to ensure that Project network folder is appropriately organized.
3) Make the original and validated health datasets "read only" so that they will not be accidentally altered.
4) Conduct routine Data Quality Assessments and report to PI and QA Manager.

8. Technical Systems Assessments

The annual Technical Systems Assessment will be conducted by the Center's QA Manager. The QC Reviewer will work with the QA Manager to provide documentation needed to facilitate this effort. The Assessment will ensure that progress is being made on the study, that the network drives are well organized and up-to-date, and that appropriate data cleaning procedures are being followed. The Assessment will ensure that the most current estimates from the air quality models are present on the Emory network drive, and that all of the ongoing analyses are using the most current estimates of air quality. The Assessment will ensure that all of the ongoing analyses have clearly documented how the data used in analyses relate to the original and validated health datasets.

9. Computer Hardware and Software

Computer hardware to be utilized for this project includes Dell brand desktop computers provided by the RSPH Information Technologies. Current generation systems are Intel Core 2 Duo processors. Standard statistical software will be used to analyze the data and present results, including SAS software, R software, Microsoft Excel software, and ArcGIS software. When necessary, analyses will make use of the RSPH cluster that is supported by the RSPH IT department. All data will be stored on password protected servers that are regularly backed up by RSPH IT department.

10. References

1. Sarnat JA, Holguin F. Asthma and air quality. Curr Opin Pulm Med 2007;13(1):63-6.
2. Bell ML, Samet JM, Dominici F. Time-series studies of particulate matter. Annual Review of Public Health 2004;25:247-280.
3. Brunekreef B, Holgate ST. Air pollution and health. Lancet 2002;360(9341):1233-42.
4. Pope CA, Dockery DW. Health effects of fine particulate air pollution: Lines that connect. Journal of the Air & Waste Management Association 2006;56(6):709-742.
5. Bell ML, Dominici F. Effect modification by community characteristics on the short-term effects of ozone exposure and mortality in 98 US communities. American Journal of Epidemiology 2008;167(8):986-997.
6. Zanobetti A, Schwartz J. Cardiovascular damage by airborne particles: are diabetics more susceptible? Epidemiology 2002;13(5):588-92.
7. Maclure M. The case-crossover design: a method for studying transient effects on the risk of acute events. Am J Epidemiol 1991;133(2):144-53.
8. Lu Y, Symons JM, Geyh AS, Zeger SL. An approach to checking case-crossover analyses based on equivalence with time-series methods. Epidemiology 2008;19(2):169-175.
9. Lu Y, Zeger SL. On the equivalence of case-crossover and time series methods in environmental epidemiology. Biostatistics 2006.