Quality Assurance Project Plan for Project 3

EPA Clean Air Research Center

Southeastern Center for Air Pollution and Epidemiology

Emory University and Georgia Institute of Technology

Management Approvals:

10/6/11

Matt Strickland
Project 3 Principal Investigator
Date

Mitchel Klein

10/10/11

Mitch Klein
Quality Assurance Advisor
Date

Guangxuan Zhu
Quality Assurance Advisor
Date

Kate Hodgins
Quality Assurance Manager
Date

10/6/11

Paige Tolbert
Co-Director
Date

Ted Russell

10/5/11

Ted Russell
Co-Director
Date

Quality Assurance Project Plan for Project 3

EPA Clean Air Research Center

Southeastern Center for Air Pollution and Epidemiology

Emory University and Georgia Institute of Technology

Management Approvals:

Matt Strickland Date
Project 3 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

1. Background/Purpose

In utero and early life experiences affect physiological development and can influence sensitivity to environmental factors throughout life. In this project we will explore the interplay between certain early life events, characterizations of air pollutant mixtures developed as part of the Center's Mixtures Characterization Toolkit, and a range of pediatric health outcomes using two large, population-based birth cohorts. Using the statewide birth cohort, we will investigate acute effects of air pollution mixtures on respiratory health outcomes and ear infections in children, and we will assess whether children who were born premature or low birth weight are more sensitive to ambient air pollutant concentrations than their counterparts. We will also investigate whether pollution levels during pregnancy are associated with the risk of preterm delivery and low birth weight. Using the Kaiser Permanente birth cohort we will examine whether pollutant mixtures during the first year of life are associated with the incidence of asthma. The findings will be important for better understanding the health effects of pollutant mixtures on susceptible populations of children.

Research questions to be evaluated:

- Are short term changes in air pollution levels associated with emergency department visits for asthma, bronchiolitis, or otitis media in children? Are these effects different among children born premature or low birth weight?
- Are air pollution levels during pregnancy associated with preterm delivery or low birth weight?
- Are long-term levels of air pollution associated with the incidence of asthma in children?

To answer the first of these questions we will obtain data on all live births from the Office of Health Indicators and Policy, Georgia Division of Community Health, and from the Georgia Hospital Association. These data have already been linked, allowing us to know whether a particular ED visit was to a child who was born preterm (< 37 weeks gestation) or low birth weight (<2500 grams). Estimates of air pollution will come from the MC Toolkit, which is described in the Air Quality Core QAAP. Five different air quality data products from the MC Toolkit will be used in the analyses:

- Community Multiscale Air Quality (CMAQ)–receptor data assimilation model for spatial interpolation of ambient monitoring data
- $PM_{2.5}$ source apportionment that incorporates a chemical mass balance approach to provide refined source profiles and contributions
- CMAQ-Mobile Matrix/CALINE grid integration for fine-scale modeling of near-roadway gradients
- Satellite remote sensing data providing regional air quality information and more detailed characterization of biomass burn events
- Measurements and modeling of Reactive Oxygen Species

The health data and the air quality data will be linked using ZIP codes and date of ED visit. Associations will be evaluated using case-crossover models. We will use confidence intervals and sensitivity analyses to evaluate whether these hypotheses are correct.

To answer the second question we will use the data on all live births from the Office of Health Indicators and Policy, Georgia Division of Community Health. Estimates from the MC toolkit will be linked with the mothers using geocodes and the date of birth. Associations will be evaluated using logistic regression. We will use confidence intervals and sensitivity analyses to evaluate whether these hypotheses are correct.

To answer the third question we will work with Kaiser Permanente (KP) on the creation of the retrospective cohort. These data reside and are owned by KP. Estimates from the MC toolkit will be linked with the children using geocodes and the date of birth. Associations will be evaluated using logistic regression. We will use confidence intervals and sensitivity analyses to evaluate whether these hypotheses are correct.

2. Design

Products to be produced include estimates from the MC Toolkit (described in Air Quality Core QAAP) and results from regression models, which will be disseminated in the peer reviewed literature and at academic conferences. A significant amount of technical work will go into the regression models, including evaluations of data quality using descriptive statistics and range checks, as well as evaluating the sensitivity of the results to alternative model specifications.

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

- Obtain the health datasets.
- Successfully estimate air pollutant concentrations (described in Air Quality Core QAPP).
- Link the air quality and health data using dates and ZIP codes.
- Conduct epidemiological analyses using appropriate statistical models, including analyses stratified by gestational age and analyses stratified by birth weight.
- Disseminate results to the scientific community through the peer reviewed literature.

Statistical models will be deemed “appropriate” if the distribution of residuals for the outcome variable, conditional on measured and modeled covariates, approximates the error distribution specified in the regression model. We will evaluate this by looking at residuals and model diagnostics.

If we successfully accomplish these tasks then we will have created statistical models that directly address our primary research questions.

3. Data Gathering Methods

In addition to the data gathering methods for the air quality data (described in the Air Quality Core QAPP), there will be three data gathering methods used to collect health data.

- Vital records data from the Georgia Division of Community Health. To obtain these data we will send the health department the Emory IRB approval, the complete HIPAA waiver, the protocol, and PHI-DUM form as required by the Office of Health Indicators and Planning.
- Emergency department data from the Georgia Hospital Association. To obtain these data we will send GHA the Emory IRB approval, the complete HIPAA waiver, the protocol, and the data use agreement as required by the Georgia Hospital Association.

- Kaiser Permanente data. To obtain these data we will send KP the Emory IRB approval, the complete HIPAA waiver. There may be issues with these data leaving KP; we will either conduct analyses on site at KP or we will work with KP to link the data there and bring a de-identified dataset back to Emory.

Once data have been obtained they will be stored on password protected network drives at Emory University that are regularly backed-up to prevent data loss. All documentation related to data collection will be retained so that these data can be reviewed later if necessary.

4. Data Quality

Data quality will be evaluated for each of the health datasets:

- For the birth data we will use descriptive statistics to examine the concordance between the birth weight and the gestational age to look for implausible combinations. If we suspect that a combination is implausible we will flag it and then conduct sensitivity analyses to evaluate whether results are sensitive to the inclusion/exclusion of these children.
- For the ED data we will use descriptive statistics examine temporal patterns in the daily counts of ED visits for the outcomes of interest. If there are sudden, abrupt changes in the counts we will examine the data closely (e.g., at the hospital-level) to evaluate whether it appears that the change might be an error. If so we will flag the event and evaluate the sensitivity of our results to the inclusion/exclusion of these visits.
- For the KP data, we will conduct descriptive statistics on all the data elements, looking for outliers and implausible values. When a data point is in question we will go back to the KP records to determine whether the value is plausible or due to human error. If it is due to human error we will whenever possible revisit the original data source and make a change to the data in the analytic dataset.

The data are *representative* because they include all individuals in the cohort. As is true with all science, the extent to which our results will be generalizable to other pediatric populations is ultimately unknowable, although we hope that our large, inclusive birth cohorts will have good external validity. We expect that the *accuracy* and *completeness* of the data will be adequate but imperfect. This is a well-recognized issue in the use of administrative health data. Measurement error should be non-differential with respect to air pollution, and the hope is that the gains in statistical power (via very large sample size) will compensate for the loss of power due to measurement error.

Data quality activities for the air quality data will predominantly be conducted by members of the air quality core. However, before using the air quality data in the epidemiological analyses for this project, we will also examine the air quality data descriptively to look for data quality issues. If there is a suspected issue we will work with Guangxuan Zhu (Air Quality QA Advisor) and Ted Russell (PI of Air Quality Core) to resolve it.

5. Data Reduction

The major data reduction technique will be regression models. We may also make data reductions by removing implausible data points, although our overall preference will be to be lenient with respect to which data points are included in the analysis.

All ongoing analyses will be able to document how the sample used in the analysis differs from the total population in the dataset. All exclusions will be clearly justified.

Short-term associations will be investigated using a case-crossover approach with subjects serving as their own controls, and the mixture characterization (from the MC Toolkit) during the event period compared with the mixture characterization during one or more referent periods. Our decision to use a case-only study design reflects our desire to limit our study’s vulnerability to confounding. For a variable to be a confounder in our proposed analyses, it must be a risk factor for emergency department visits that varies in a systematic manner with temporal changes in ambient pollutant mixtures. Individual-level risk factors, such as exposure to environmental tobacco smoke or socioeconomic status, are not associated with short-term changes in air pollutant concentrations and are therefore not confounders. However, meteorological changes, holidays, day-of-week patterns, and back-to-school effects do vary temporally and will be thoroughly assessed as potential confounders. Because temperature is a strong predictor of both air pollutant concentrations and respiratory infections, referent periods for the case-crossover analysis will be selected by matching on same-day temperature within calendar month. A general form of the conditional logistic regression models used for these analyses is:

$$logit\ (Y_{ij}) = \beta_0 + \beta_1 (MC_{ij}) + \theta\ (temporally\text{-}varying\ covariates_{ij})$$

where Y_{ij} is the dichotomous outcome indicating whether or not subject *i* visited the emergency room on day *j*; β_1 represents the effect of the mixture characterization (MC) during the exposure period for subject *i* on day *j*; and θ is a vector representing the effects of potential confounders, such as the day of week. The MC_{ij} assigned to subject *i* on day *j* is based on admission date and residential zip code and can correspond to measured or modeled pollutant concentration(s), pollutant source contributions, or any other output from the MC Toolkit. Various approaches (e.g., stratification, interaction terms) will be utilized to investigate possible effect modification by gestational age or low birth weight. A full exploration of pollutant lag effects, including constrained distributed lag models, will be conducted for asthma, bronchiolitis, and otitis media. Modeling approaches for characterizing concentration-response relationships will include categorization by quintiles and use of loess smoothers in generalized additive models.

Preterm birth, defined as birth before 37 weeks’ gestation and calculated from the first day of the reported last menstrual period date, will be analyzed as a dichotomous outcome. A general form of the model to assess associations between air pollution mixtures and preterm birth is:

$$logit\ (Y_i) = \beta_0 + \beta_1 (MC_{ij}) + \delta (individual\text{-}\ level\ covariates_i) \qquad (2)$$

where Y_i is the dichotomous outcome indicating whether or not subject *i* was born preterm; β_1 is the effect of exposure to the pollutant mix during the pregnancy window of interest; and δ is a vector of effects for the individual-level covariates for subject *i* such as maternal education and race. For preterm birth we will focus on exposures during late pregnancy, based on our previous findings, and the hypothesis that inflammatory insults in late pregnancy can trigger a cascade of events leading to the initiation of early parturition. However, because little is known about the relevant exposure period for preterm delivery, we will assess other exposure periods during pregnancy, including early pregnancy when implantation and placentation occur.

Birth weight, in grams, will be analyzed as a continuous variable and adjusted for gestational age. A general form of the model for the continuous birth weight analysis is:

$$E(Y_i) = \beta_0 + \beta_1 (MC_{ij}) + \gamma (gestational\ age) + \delta (other\ individual\text{-}\ level\ covariates_i) \quad (3)$$

where Y_i is birth weight in grams of subject *i*; β_1 is the effect of exposure to the pollutant mix during the pregnancy window of interest; γ is a vector of effects for gestational week (modeled as indicator variables to account for the nonlinear effect of gestational age on birth weight); and δ is a vector of effects of the other individual-level covariates for subject *i*. For the birth weight analysis we will investigate exposures throughout pregnancy, including a novel application of constrained distributed lag models where each "lag" represents exposure during one of the nine months of pregnancy; although limited to full-term infants, this analysis may help to identify specific periods of vulnerability during pregnancy, a major challenge in this area of research.

We will also investigate associations between air pollutant mixtures during the first year of life and incident asthma in childhood. A child who meets at least one of three criteria will be considered asthmatic: (1) one inpatient visit with a primary diagnosis of asthma, (2) two outpatient visits for asthma, or (3) one outpatient visit for asthma plus one prescription for an asthma medication. To be conservative, our primary analysis will ignore asthma diagnoses made before age five, since diagnosis of asthma in young children is challenging, and the majority of toddlers who present with persistent wheeze outgrow the condition. Sensitivity analyses will be conducted to explore younger age cutoffs (e.g., asthma diagnoses after age 3 years) and alternative asthma definitions (i.e., characterizing the presence or absence of asthma using various combinations of inpatient visits, outpatient visits, and prescription history). Although it is substantially smaller than the statewide cohort, the detailed information on asthma diagnoses available for children in the Kaiser Permanente birth cohort is a key advantage, and the "closed" nature of the cohort enables us to create Cox proportional hazards models to investigate associations between air pollutant mixtures during the first year of life and incident asthma. The general form of the stratified Cox proportional hazards regression model is:

$$h_g(t,X)= h_g(t)\ exp[\beta_0 + \beta_1 (MC_{ij}) + \delta (individual\text{- } level\ covariates_i)] \qquad (4)$$

In this model, each subject either has the event (an asthma diagnosis) or is censored (i.e., follow-up ends or the subject leaves the HMO). In the above model *t* is the person*time at risk; β_1 is the effect of exposure to the pollutant mix during subject *i*'s first year of life; and δ is a vector of effects of individual-level covariates for subject *i*. The subscript *g* indicates a stratified Cox proportional hazards model. For our primary analyses, we will stratify the cohort into small (e.g., 3-month) time intervals according to the subjects' birth dates. By forming several "mini-cohorts" of children born at approximately the same time we hope to account for confounding by time-varying risk factors that may not have been measured. Air pollutant mixtures will be assigned based on the child's residence during the first year of life. If the child changes residence during infancy, a weighted average exposure value will be calculated based on the proportion of the year spent at each residence. Interaction by preterm birth or low birth weight can also be assessed in this cohort, although these analyses may have limited power.

6. Interaction of the Players

The PI of the project is Matt Strickland. He will direct the project and has the authority to change how the various players interact. He is supported by Co-PIs from Emory (Lance Waller) and from Georgia Tech (Ted Russell). Since Dr. Waller and Dr. Russell are the PIs of the Biostatistics Core and the Air Quality Core (respectively), they will facilitate interaction between this project and the Cores. The efforts of the project will be overseen by the Core's two QA Advisors, Guangxuan Zhu and Mitch Klein, who will work with the PI to conduct annual reviews of project methods and QA procedures. Craig Hansen from Kaiser Permanente will direct the creation of the KP birth cohort and will be a part of the epidemiological analyses based on those data. Randy Guensler, Jim Mulholland, Lyndsey Darrow, Yang Liu, and Paige Tolbert are co-investigators on the project. All investigators with access to health data will maintain their CITI certification by taking refresher courses every-other-year (or as specified by Emory University). These will be stored on the network drive and the annual IRB renewal will not be awarded until all CITI tests from study personnel are up to date.

For the statewide birth cohort Katie Gass will be the QC Reviewer for the project and will keep records of QA/QC procedures and will document any deficiencies as they occur. For the KP birth cohort Audrey Flak will serve as the QC Reviewer once data collection activities start. These two QC Reviewers will send these documents to Kate Hodgins, the Center's QA Manager.

7. Technical Systems Assessments

The annual Technical Systems Assessment will be conducted by the Center's QA Manager. The two QC reviewers 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 sample used in that analysis relates to the total population in the database.

8. Extramural Agreements and Contracts

Extramural agreements and contracts include data use agreements with the Office of Health Indicators and Policy (OHIP), Georgia Division of Community Health, and with the Georgia Hospital Association. For OHIP we will complete the Data Use Policy Form (OHIP form 003 ver2.2). For the Georgia Hospital Association we will complete the Data Use and License Agreement (revision 09/2010 to address IRB waiver). This contract with the Georgia Hospital Association will be routed to Emory's Office of Grants and Contract before the data are purchased.

9. Data Management

Health data will be stored at Emory on password protected network drives that are regularly backed-up. 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. Final health datasets will be stored as "read only" on the network drive and then modifications can be made to copies of these datasets as needed. The tasks of the QC Reviewers relate to data management, specifically:

1) Check to ensure that our Emory network drive has the most current versions of all the different air quality models that are on the Georgia Tech website.
2) Check to ensure that Project network folder is appropriately organized.
3) Make the full cohort datasets read only so that they can't accidentally be altered.
4) Describe the full cohorts quantitatively (before any exclusions) (e.g., sample size, ages, etc).
5) Ensure that all ongoing analyses can clearly document how the sample used in the analysis relates to the number of people in the full cohort (i.e., exclusions are ok, but they need to be documented)

10. Computer Hardware and Software

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