Core B:
Biostatistics Core

Director: Brent Coull
Co-Directors: Francesca Dominici, Joel Schwartz
ABSTRACT – Biostatistics Core (Core B)

The Biostatistics Core will provide centralized statistical and analytical expertise to all Center projects. Core Faculty are drawn from the Environmental Statistics Program within the Department of Biostatistics and the Exposure, Epidemiology, and Risk Program within the Department of Environmental Health at the Harvard School of Public Health. Core members provide expertise in the general statistical methods needed for the projects, such as linear regression and ANOVA, correlated data analysis (including longitudinal and spatial data analysis), measurement error, semiparametric and nonparametric (smoothing) models, meta-analysis, structural equation models, and Bayesian data analysis.

The Biostatistics Core will provide: 1) support for statistical analysis for all five proposed Projects, including, design consultation and analytical work; and 2) training for investigators in statistical issues needed for the data analysis and in SAS and R software. In addition, Core activities will include mission-related methodological research to develop needed statistical tools when existing methodology does not fully address the scientific question of interest.

Power and sample size calculations are critical components in all the Projects. Prospective calculations will ensure that Center project designs afford high power to detect meaningful differences. Core investigators have worked closely with Center investigators to determine effect sizes of interest and to calculate the numbers of samples necessary to achieve a desired level of power, usually 80% and 90%.

To the extent allowable by design and outcome commonalities among the five Projects, Core investigators will ensure that a unified approach to modeling strategy and choice of data transformations is applied to all Center data. Statistical analyses will apply appropriate exploratory data analysis techniques, such as univariate explorations of the data, distributional checks, and outlier identification to data from all projects. Residual analysis and other model diagnostics will be routinely used to confirm model fit, identify possible nonlinear relationships between predictors and outcomes, and identify highly influential data points. Once the data have been checked and modeling assumptions verified, primary analysis methods will include ANOVA and regression techniques, with the particular form of the outcome and correlation structure of the data dictating the particular method. The main analytical methods will be linear models/generali zed linear models, semi-parametric regression modeling (smoothing), mixed/multivariate models for correlated responses, growth curve modeling, and Bayesian hierarchical models.

Development of new statistical methods (including software development and dissemination) will play an important role in the proposed Center Projects. These methods include spatio-temporal models for estimating exposures to PM, methods to address model uncertainty, novel distributed lag models, and exposure measurement error corrections. Future methodological development will focus on additional spatio-temporal modeling of pollution, the development of clustering methods to characterize exposure to complex pollution mixtures, and diagnostic tools to detect confounding.
1. SPECIFIC AIMS

The Biostatistics Core will provide centralized statistical and analytical expertise to all projects. Core Faculty members are drawn from the Environmental Statistics Program within the Department of Biostatistics and the Exposure, Epidemiology, and Risk Program within the Department of Environmental Health at the HSPH. Core members provide expertise in the general statistical methods needed for the Projects, such as linear regression and ANOVA, correlated data analysis (including longitudinal and spatial data analysis), measurement error, generalized additive models and semi-parametric (smoothing) methods, meta-analysis, structural equation models, and Bayesian data analysis. Aims of the Biostatistics Core are to provide: 1) support for statistical analysis for all five proposed Projects, including substantial design consultation and analytical work and; 2) training for graduate students and investigators in both statistical issues involved in the data analysis as well as in SAS and R software. In addition, the Core will conduct mission-related methodological research to develop needed methods and software when existing methodology does not fully address the scientific question of interest. Important Core initiatives for the proposed Center include spatial and spatio-temporal modeling of air pollution exposures (Sections 3.2.1, 4.4.1, and 4.4.3), the development of a unified framework for assessing the health effects of pollutant mixtures (Sections 4.2.1 and 4.4.2), and the application of flexible regression methods for estimating the effects of air pollution on cognitive and neuropsychological function, vascular and endothelial function, and somatic growth (Section 4.2.2).

2. INTRODUCTION

As part of our existing Harvard PM Center, Core investigators Brent Coull, Joel Schwartz, and Antonella Zanobetti provide statistical support, conduct novel data analyses, and address methodological issues arising from new investigations. In addition, we are fortunate to have Dr. Francesca Dominici joining us as Co-Director of the Core. Dr. Dominici is an international leader in the development and application of biostatistical and epidemiological methods to assess the health effects of air pollution. Prior to coming to Harvard she served as the Director of the Johns Hopkins PM Center. A major strength of the Core is its strong connection to the Environmental Statistics Program in the Department of Biostatistics at the HSPH, which is co-directed by Drs. Coull and Dominici. This allows the Center to leverage additional personnel time via an existing NIEHS-funded T32 Training Program in Environmental Statistics, which supports 8 predoctoral fellows and 2 postdoctoral fellows and provides a rich pool of funded trainees who work on Center-related research. In turn, the Center provides a rich set of air pollution-related problems that engage trainees interested in Environmental Statistics and Environmental Epidemiology. This synergistic arrangement has benefited the current and previous Harvard PM Centers. During the current award period, 5 predoctoral fellows, 6 postdoctoral fellows, and one Assistant Professor from the HSPH Biostatistics Department have contributed to Center-related projects. Such leveraging of training-related effort enables the Core to meet the analytic needs of the Center with limited resources.
3. PROGRESS REPORT

Activities of the Biostatistics Core during the current the Harvard PM Center include:

3.1 Data Analysis: Because an important goal of the existing Center has been to investigate associations between biologic responses and exposures, our main thrust of data analysis has been the use and extensions of regression techniques. However, because the nature of the outcomes, study design, and exposure characterization vary by study, analytical approaches differ by project. Following is a short description of the Core’s role in our current PM Center Projects:

3.1.1. Normative Aging Study: We used regression-based methods to estimate associations between outcomes and PM exposures. Specifically, due to the longitudinal design of the health assessments, we used hierarchical models with random effects to account for within-subject repeated measures.\textsuperscript{1,2} Specifically, let $Y_{ij}$ be the response from subject i on day j. For concreteness, we consider an NAS analysis involving time-varying covariates (pollution and temperature) and individual-level covariates (diabetes and age at baseline). We have relied heavily on hierarchical models of the form:

$$Y_{ij} = \beta_0 + \beta_1 pollution_{ij} + \beta_2 age_i + \beta_3 diabetes_i + \beta_4 temp_{ij} + \ldots + u_i + \epsilon_{ij} \quad (1)$$

We typically include day of the year as an additional confounder in order to correct for seasonal trends in the outcome. Here $u_i$ is a subject-specific intercept that reflects unexplained heterogeneity in subjects’ overall outcome levels. We make the standard assumption that the $u_i$ are generated from a normal distribution with variance estimated from the data, and use diagnostic methods we developed to check this assumption.\textsuperscript{3-5} Such models for continuous outcomes can be easily fitted using standard software, such as PROC MIXED in SAS and the lme() function in R. For designs having moderate-to-large numbers of repeated measures, we also consider models containing random slopes that allow for heterogeneity in the health effects of air pollution across subjects. Assessment of measured factors that affect susceptibility is readily incorporated into this framework by adding (pollution x factor) interaction terms into the model. Inclusion of these interaction terms can be viewed as covariate information that can explain the variation in the random slopes. Important extensions of model (1) that we used in our NAS analyses included generalized linear mixed models (GLMMs) and generalized estimation equations (GEEs) for discrete outcomes.

3.1.2. Human and Animal CAPs Exposure Studies: Both animal and human toxicological experiments have used partially randomized designs, in the sense that exposure (yes/no) is randomly assigned but the composition of the assigned exposure is random. That is, subjects were randomized to receive either control (particle-free air) or concentrated ambient particles (CAPs) exposures, either within straightforward factorial structure (early/late morning animal exposure studies) or more complex crossover designs. On a given exposure day, the Harvard Ambient Particulate Concentrator (HAPC) produces essentially random CAPs exposures, in terms of PM mass and composition. This presents statistical challenges exceeding those encountered in classic controlled-exposure studies. To analyze these data, we have conducted multiple analyses that use exposure metrics of increasing sensitivity. First we used (potentially multi-way) ANOVA techniques that treat exposure as a treatment, to assess overall differences between CAPs and filtered air responses (i.e., a binary exposure covariate). Second, to assess univariate associations between mass levels or composition of CAPs and health outcomes, we
conducted single-pollutant analyses in which a separate regression model is fitted using the biologic response as the dependent variable, and either mass, particle number, or a single elemental concentration as the exposure metric. Third, to confirm univariate findings, we conducted multiple pollutant analyses to investigate the combined effects of pollution sources. We have thereby linked biological outcomes to predictors including: 1) multiple tracer elements of previously-defined pollution sources; and 2) factor scores obtained from source-apportionment factor analysis to represent PM source contributions. Such random experimental exposures are typically nested within a complex design (e.g., crossover or repeated-measures). Thus, although we refer to these three levels of analysis as “ANOVA” and “regression”, our strategy of increasing the sensitivity of the exposure metric is usually nested within regression extensions that respect the study design and the correlation structure of the data.

3.1.3 TERESA Study: In these animal exposure studies, rats were randomly divided into treatment groups, each receiving a given exposure scenario. The protocol was also replicated in different rat models (healthy, spontaneously hypertensive, and myocardial infarction rats). Outcomes of interest included chemiluminescence, bronchoalveolar lavage (BAL), BUXCO pulmonary measures, blood parameters, and EKG recordings. Data collection in rats is still underway, but we have been heavily involved in analyzing an earlier TERESA study focusing on power plant emissions. Briefly, we used multi-way ANOVA methods to assess differences between groups defined by exposure scenario and power plant. Also, we embedded ANOVA-based concepts into a smooth-trend model for continuously-measured respiratory endpoints. Although the primary level of analysis was power plant and exposure scenario, there was also variability in individual pollutant levels across experimental runs within scenarios. We used regression and regression tree-based methods to assess univariate and multivariate associations between component levels.

3.1.4 Additional Analytic Approaches

Additive Modeling: A primary objective of the Centers RFA is the assessment of the exposure/concentration-response relationships for PM and/or ozone and specific health outcomes. We have used additive extensions of the regression models outlined above to flexibly estimate the functional form of these relationships from the data. For instance, an extension of model (1) that specifies an arbitrary smooth exposure-response function is:

\[
Y_{ij} = \beta_0 + f(pollution_{ij}) + \beta_{2age} + \beta_{3diabetes} + \beta_{4temp_{ij}} + \ldots + u_i + \varepsilon_{ij}
\]  

(2)

where the form of \(f\) is estimated from the data. Such models can be routinely fit using the \texttt{gamm} function in the R software package, with the smoothing parameter governing the smoothness of \(f\) chosen by generalized cross-validation or restricted maximum likelihood (REML). We also use this formulation to flexibly control for confounders as well. For instance, we replace \(\beta_{4temp_{ij}}\) with \(g(temp_{ij})\).

Case–Crossover Analyses: The case-crossover design is a method for estimating the risk of an acute outcome associated with short-term exposure to a given pollutant. In a case–control study, each person who experienced an event is matched with herself/himself at a nearby time period during which s/he did not have the event. The subject’s characteristics and exposures at the time of event are compared with control periods during which such an event did not occur. During the current Center, we applied such case-crossover analysis to analyze the association between estimated BC levels and mortality. We have also collaborated with Dr. Mittleman (PI,
Framingham Project, Project 3), to investigate the impact of misclassification of onset time on 
estimated health effects.\textsuperscript{14}

**Time to Event Data:** Events such as cardiovascular deaths are treated as censored survival time 
data using standard Cox regression models. To allow for the clustering of observations by 
location, we use the random effects Cox model.\textsuperscript{15} Methods for additive extensions of the Cox 
model have been used to allow for non-linear covariate effects. These additive Cox models can 
be implemented using the pspline function in the R software package.

**3.2 Methodological Development:** During the current Center we developed new statistical methods 
when existing methods did not fully address the scientific question of interest.

**3.2.1 Spatio-temporal Models:** We developed, validated, and published a spatio-temporal model 
predicting 24-hr average measures of BC at a given address.\textsuperscript{16} The dataset included >6,021 
observations from >2,079 unique exposure days at 82 monitoring sites. Predictions of BC levels 
were computed accounting for meteorological and other characteristics (e.g., weekday/weekend) 
of a particular day and measures of land use (e.g., cumulative traffic density within 100 meters, 
population density, distance to nearest major roadway, and \%urbanization) at a given location. 
The model allowed for space-time interactions in BC concentrations by including interaction 
terms between the temporal meteorological predictors and source-based geographic variables. 
We used regression splines to allow main effect terms to nonlinearly affect exposure levels, and 
thin-plate splines, a 2-dimensional extension of regression splines, to model additional spatial 
variability unaccounted for after including all relevant spatial predictors in the model. Separate 
models were fit for warm and cold seasons. The prediction equation used concentration data 
from the Boston Supersite as a predictor to reflect average pollutant concentrations for a given 
day. Predicted daily concentration levels showed a >3-fold range of variation in exposure across 
measurement sites (adjusted $R^2 = 0.83$). We later conducted monthly monitoring campaigns at 30 
additional sites as a validation sample. The out-of-sample predicted $R^2$ based on this validation 
sample was 0.53, compared to an $R^2$ of 0.10 for the approach that uses only the central-site 
monitor to predict location-specific BC levels. Further refinements of the model are on-going 
(Section 4.4.1). We used the BC exposure predictions in several Center-supported studies 
assessing the association between health outcomes and predicted pollutant levels in Eastern MA, 
including lung function\textsuperscript{17}, cognitive function\textsuperscript{18}, birthweight\textsuperscript{19}, telomere length\textsuperscript{20} and blood 
pressure\textsuperscript{21} outcomes. We outline how exposure estimates from this model will be used in the new 
Center in Section 4.2.

**3.2.2 Measurement Error:** We developed statistical methods that correctly account for the 
uncertainty associated with predictions from exposure models such as described above when 
such predictions are used as covariates in health effects analyses.\textsuperscript{19} This work provided a 
measurement error framework for this setting, showing that spatial smoothing induces a 
Berkson-type measurement error with non-diagonal error structure. Using our proposed methods, 
we estimated the association between traffic-related PM exposures and birthweights in the 
greater Boston area. Other work developed Bayesian measurement error adjustments for multi-
city time series studies.\textsuperscript{22,23} Our novel proposed methods do not require knowledge of the 
measurement error magnitude, and produce estimators much more efficient than those from 
analogous two-stage approaches we proposed earlier.\textsuperscript{24}

**3.2.3 Statistical Methods for Assessing Source-Specific Health Effects of Air Pollution:** We 
investigated the use of structural equation models for assessing source-specific health effects of
air pollution. This approach corresponds to fitting a source apportionment (i.e., factor analysis) model and the health outcome model jointly, such that inferences on the health effects account for the fact that uncertainty is associated with the estimated source contributions. Using simulation, we compared the performance of this approach and two prior ones, including a tracer approach and a two-stage source apportionment approach. We illustrated differences between the two methods by applying them to CAPs data generated by our group. We extended this research to assess how sensitive effect estimates obtained from a source apportionment/structural equations model fit are to misspecification of the model. We found that receptor models that: 1) maintained the non-negativity of chemical concentrations with a multiplicative error structure; and 2) also incorporated meteorological information and an adjustment for temporal correlation in the exposures, generally produce larger, more significant estimated effects.

### 3.2.4 Distributed Lag Modeling

We estimated the temporal pattern of association between a health outcome and previous exposure using distributed lag models that estimate a smooth function of the association between outcome and exposure lag. We used distributed lag models to estimate mortality displacement in the association of ozone with mortality. We also developed new formulations of distributed lag models to deal with data complications arising specifically in PM epidemiology. Specifically, we proposed wavelet-based distributed lag models as a measurement error correction method and new distributed lag models for time-series mortality studies.

### 3.2.5 Novel Methods for Addressing Model Uncertainty

One criticism of the existing PM epidemiologic analyses is that multiple sources of uncertainty are involved in obtaining health effect estimates. Our group has argued that the popular approach to model averaging is not appropriate for uncertainty concerning confounder adjustment, and proposed a new framework. We also showed that model averaging is an effective approach for estimating both the functional form of dose-response relationships and the distributed lag structure of the relationship between an outcome and an exposure.

### 4. RESEARCH METHODS:

#### 4.1. Study Design

To ensure that all Center projects will have sufficient power to detect meaningful differences, we perform prospective power and sample size calculations. We have worked closely with Center Investigators to: 1) determine effect sizes of interest; and 2) calculate the number of samples necessary to achieve a desired level of power, usually 80% and 90%. Power and sample size calculations for the proposed experiments were presented in the respective Project proposals. For input into the power calculations, we use estimates of the within-person and between-person variability obtained from previous studies for each endpoint, most of which were conducted by Center Investigators. We also use existing data to estimate variability of exposure. All power calculations are based on two-sided tests at the 0.05 level.

#### 4.2. Data Analysis

We will apply exploratory data analysis techniques including univariate explorations of the data, distributional checks, and outlier identification to all Project data. For model building, we will use residual analysis and other diagnostics (such as semi-variograms for spatial data) to confirm model fit, identify possible nonlinear relationships between predictors and outcomes, and identify highly influential data points. After data have been checked and modeling assumptions verified, we will apply ANOVA and regression techniques. The particular form and correlation structure of the data will dictate the method. As in prior work, our main
analysis methods will be linear models/generalized linear models, multi-way ANOVA, semi-parametric modeling, and mixed/multivariate/latent variable models for correlated responses. We anticipate that statistical tools developed within our mission-related statistical research (Section 4.3) will become staples of our toolbox for analyzing Center data.

4.2.1. Analytic Strategies for Specific Exposure Hypotheses: In the Center's Cohort Studies (NAS, Framingham and Viva) and in the National Study, primary hypotheses concern the effects of short-term and intermediate/long-term exposures. As outlined in the Exposure Core, we will assess effects of 6 different exposure metrics: 1) Short-term exposures to individual pollutants; 2) Short-term exposures to sources; 3) Short-term exposures to mixtures; 4) Long-term exposures to individual pollutants; 5) Long-term exposures to individual sources; and 6) Long-term exposures to mixtures. To demonstrate how we will use these metrics to analyze health effects, we consider how they will be used in a typical regression setting. The specific form of the model to be used in a given analysis will depend on the specific study design, but for concreteness we consider each exposure metric in the hierarchical model (1):

$$\begin{align*}
Y_{ij} &= \beta_0 + [pollution]_{ij} + confounders + u_i + \epsilon_{ij} \\
\end{align*}$$

where the pollution component consists of one or more regression terms dictated by the exposure hypothesis. Briefly the form of covariate for each type of scientific hypothesis listed above is:

(a) Short-term exposures to individual pollutants: For this exposure hypothesis, the term consists of the daily (or a multi-day average) concentration of a given pollutant. For PM$_{2.5}$, O$_3$ and BC, we will use both daily Boston Supersite measurements and estimates of participant home-specific exposures. The estimated exposures will be generated as follows: 1) PM$_{2.5}$ exposures will be estimated using MODIS and GOES satellite data starting in 2001; 2) BC and O$_3$ exposures will be estimated using spatio-temporal models starting in 1995 (see Sections 3.2.1 and 4.4.1). For these pollutants, health effects analyses will follow a two-tiered approach. The first tier of analysis will use daily measures recorded at the Boston Supersite and the second tier will use the address-specific estimated exposures. The two-tiered approach allows us to combine the strengths of the two forms of data, as the central-site measurements: 1) yield continuous measures of pollution levels, thus yielding the finest temporal resolution; and 2) do not rely on modeling assumptions (and are hence less susceptible to misspecification of the exposure models). In contrast, the address-specific estimates will reduce the potential for measurement error incurred in the central site readings of local-source pollutants, which should yield more precise health effect estimates due to the Berkson nature of the error. For pollutants other than PM$_{2.5}$, BC, and O$_3$, we will use data collected at the Supersite in all health effects analyses.

(b) Short-term Exposures to individual Sources: As outlined in the Exposure Core, estimates of source-specific exposures will be constructed using source apportionment methods. For concreteness, let n be the number of days, p be the number of elements, and q be the number of hypothesized sources. If we let X denote an n x p matrix of elemental concentrations, then the source-apportionment factor analysis model is:

$$X = SP + \epsilon$$

where S is an n x q matrix of unknown source contributions and P is a q x p matrix of unknown factor loadings relating each element to each source. This model is not uniquely identified, as an infinite number of combinations for S and P yield the same model for X. Several approaches can
be used to address this non-uniqueness. We will use positive matrix factorization (PMF).\textsuperscript{35}

Assuming five identified sources, the resulting health effects model becomes:

$$\text{[pollution]}_{ij} = \beta_{1s_{1ij}} + \beta_{2s_{2ij}} + \beta_{3s_{3ij}} + \beta_{4s_{4ij}} + \beta_{5s_{5ij}}$$

(5)

where the covariates are the estimates of the source contributions contained in the row of S corresponding to day j. As a sensitivity check on this primary analysis, we will also use a targeted factor analysis within a structural equation model framework. This approach\textsuperscript{25-26} rotates the estimated factor loadings towards a solution defined by known tracer elements for each source, and fits the factor-analytic and health effects model jointly to properly account for uncertainty associated with the estimated source contribution covariates. These strategies will be applied to data from all three Center Cohort studies (Projects 2-4), as well as to data from 10 selected cities collected within the National Study (Project 5).

(c) Short-term Exposures to individual Mixtures: We propose a new clustering paradigm for estimating the joint effects of the entire pollutant mixture. Independent effects of individual pollutants within mixtures are often effectively unobservable due to highly collinear components. In this section we describe our approach for estimating a health risk associated with short-term exposure to mixtures, when this risk might be non-additive (that is, it differs from the sum of estimated risks associated with exposure to single pollutants). More specifically, first, we will use a temporal clustering algorithm (see Exposure Core) to create groups of days that have similar multi-pollutant profiles. Second, we will estimate a different health risk associated with short-term exposure to PM$_{2.5}$ total mass separately within each cluster. Under this approach, we will be able to test whether these cluster-specific health risks vary across clusters, thereby assessing the relative toxicities of the corresponding profiles of the mixture. Specifically, we will fit the following regression model:

$$\text{[pollution]}_{ij} = \gamma_{c_{1j}} + ... + \gamma_{C_{-1}c_{C-1j}} + \beta_{1c_{1j}}\text{pm}2.5_{ij} + ... + \beta_{Cc_{Cj}}\text{pm}2.5_{ij}$$

(6)

where $c_{kj}$ is an indicator variable equal to 1 if day j belongs to the cluster k, and to 0 otherwise. This model assumes that both the overall level of the health outcome and the association between PM$_{2.5}$ and the outcome varies among clusters of days having similar pollution mixture profiles. We focus on estimation of health risks associated with short-term exposure to PM$_{2.5}$ total mass due to the fact that information on other components are used to inform the classification of the pollution mixtures. We will test whether the health risks associated with short-term exposure to PM$_{2.5}$ mass vary across clusters (i.e. $\beta_1 = \beta_2 = ... = \beta_C$), and, if not, whether the health outcomes overall vary across clusters (i.e. $\gamma_1 = \gamma_2 = ... = \gamma_{C-1} = 0$).

(d) Long-term Exposures to Individual Pollutants: For hypotheses that concern the health effects associated with long-term (greater than 6 months) exposure to individual pollutants, the “pollution” term in model (3) represents an estimated address-specific long-term average concentration of a given pollutant. For PM$_{2.5}$, BC, and Ozone, we will average daily exposure estimates obtained from the data sources described in (a) above. For all other pollutants, we will apply the long-term spatial models (described in Section 4.4.3) to the long-term monitoring data to be collected during 2010-2014 as part of this Center (see the Exposure Core for a description of this proposed monitoring effort). We will validate these models using out-of-sample validation methods.\textsuperscript{19} For intermediate-term exposures, such as trimester-average and 6-month exposures needed for Viva (Project 2) analyses, preliminary data presented in the Exposure Core
suggest that averages over these timeframes of daily Boston Supersite data exhibit sufficient temporal variability for use in health effect analyses.

(e) Long-term Exposures to Sources: We will estimate long-term exposures to sources using a two-stage process. First, we will conduct separate source apportionment analyses for all of the spatial sites (see the Exposure Core for a description of the available ambient monitoring data on PM composition) in the Region using the EPA Positive Matrix Factorization (PMF) method. In order to allow factor loadings [i.e. \( P \) in equation (3)] to vary over time, this analysis will be performed separately by year. Putting together the resulting daily source contribution estimates from all years, we will then stratify these daily estimates by season. At the second stage, for each source in each season, we will spatially smooth these estimated source contributions using the long-term spatial models that are proposed in Section 4.4.3. Although we do not directly observe measures of source contributions, we will check the ability of these long-term spatial models to predict source-specific exposures using tracer elements of each source. We will do this by first constructing estimates of the concentrations of tracer elements, for locations for which we have data during 1999-2009, and comparing them to the observed values. This strategy represents an out-of-sample validation step for the tracer elements. We will use a similar strategy for data from 10 selected cities collected as part of the National Study (Project 5).

(f) Long-term exposures to individual mixtures: We will investigate whether health effects associated with long term exposure to individual pollutants differ across groups of geographical locations (e.g. counties) that have a similar air pollution mixture profile. Specifically, consider a data structure in which multivariate exposure data for several individual pollutants are available in many geographical locations. We will apply the clustering methods outlined in Section 4.4.2 and in the Exposure Core to the area-level exposure data to define groups of geographical locations having similar profiles of long-term air pollution. We will fit the following regression model:

\[
[pollution]_{ij} = \gamma_1 c_{1i} + \ldots + \gamma_{C-1} c_{C-1,i} + \beta_{i} pm2.5_{ij} + \ldots + \beta_{C} c_{C,i} pm2.5_{ij} \tag{7}
\]

where \( c_{ki} \) denotes a variable that indicates whether subject \( i \) resides in an area that belongs to cluster \( k \). For our Region-based Cohort Studies (Project 2-4), we will use a three-step procedure. First, we will apply the long-term spatial models described in Section 4.4.3 to the long-term monitoring data proposed to be collected during 2010-2014 as part of this Center (see the Exposure Core for a description of this proposed monitoring effort). Second, we will use the resulting model fits to compute estimated average concentrations for each modeled air pollution component at the zip-code level. Third, we will apply the clustering methods outlined in Section 4.4.2 to the resulting zip-code level concentration averages. For the National Study (Project 5), we will cluster counties based on their pollution mixtures, allowing us investigate effect heterogeneity by pollution mixture characteristics, rather than simply by geographic region. We will then compare the variation of the PM effects stratified by cluster group to that exhibited by traditional geographical stratification.

4.2.2 Data Analytic Methods for Center Projects: The exposure hypotheses framework (4.2.1) will be applied to analyze data for the five Center Projects as follows.

4.2.2.1 Project 1 (Toxicological Study): We will employ longitudinal methods to estimate associations between exposure and biologic parameters in the proposed cross-over design. On each exposure day, separate animal groups will be exposed to a) Sham (filtered air); b) Control
Exposure (e.g. unreacted CAPs), and; c) Exposures (e.g. aged CAPS) (see Project 1). Briefly, data from each animal are collected during two baselines (one at the beginning and the end of the experiment) and 5 exposures. Let \( Y_{ij} \) be a biologic endpoint recorded for animal \( i \) at time \( j \). We consider models of the form:

\[
Y_{ij} = \beta_0 + \beta_1 \text{expo}_{ij} + \beta_2 \text{cexpo}_{ij} + \beta_3 j + u_i + \epsilon_{ij},
\tag{8}
\]

where \( \text{expo}_{ij} \) and \( \text{cexpo}_{ij} \) are indicator variables equal to 1 if subject \( i \) is subjected to the exposure or controlled exposure, at time \( j \) and \( \beta_3 j \) represents a linear term corresponding to exposure occasion (1-9). We will consider interaction terms that allow the effect of exposure to vary over time. If necessary we will also consider more general forms for the effect of time and more flexible correlation structures than implied by random subject-specific intercepts. We have used this class of models to analyze similar microsphere data in a study of CAPs effects on coronary ischemia in dogs.\(^{36}\) We will extend the above model to use concentrations of individual pollutants or multiple pollutants, depending on the hypothesis of interest. In a single pollutant analysis, let \( \text{conc}_{ij} \) denote the concentration for that pollutant of interest for animal \( i \) at exposure occasion \( j \). The model becomes:

\[
Y_{ij} = \beta_0 + \beta_1 \text{conc}_{ij} \text{expo}_{ij} + \beta_2 \text{conc}_{ij} \text{cexpo}_{ij} + \beta_3 j + u_i + \epsilon_{ij},
\tag{9}
\]

The model estimates an exposure-response slope under both exposure and controlled exposures, and yields a test of whether these slopes are equal. This test will assess the evidence that exposure increases the toxicity of particular air pollution components. Extensions to multi-component models are straightforward.

4.2.2.2 Project 2 (NAS Cohort): Our analyses of NAS data will continue to use linear regression, semi-parametric regression, Cox proportional hazard models, and longitudinal models. Responses of interest will include measures of vascular and endothelial function, inflammation, and cognition. We will investigate factors affecting susceptibility/vulnerability in the above models by including interaction terms for pollution and an indicator of potential modifier of interest. We will consider spatial regression models to account for potential spatial correlations among outcomes for subjects located close together in space. We will continue to use and develop spatial-temporal and long-term spatial models to refine estimates of exposure for pollutants exhibiting spatial variability (see Sections 3.2.1 and 4.4.1). We will control for confounding by pre-filtering exposure using time-series methods.\(^{37}\) This will eliminate potential confounding by seasonality and temperature. This method represents a form of G-estimation of health effects in a causal inference framework.\(^{38,39}\) Additional work by our group includes that described in Dominici et al.\(^{40}\) and Peng et al.\(^{41}\)

4.2.2.3 Project 3 (Framingham Cohort): Our analyses of Framingham Offspring and Gen 3 data will use linear regression, semi-parametric regression, Cox proportional hazard, and longitudinal models. Responses of interest will include cognitive outcomes, such as the Mini-Mental Status Exam (MMSE), and markers of vascular and endothelial function. Models for the vascular and endothelial parameters will be similar to those used in the NAS cohort (Section 3.1.1). For longitudinal MMSE measures, we will use a nonparametric regression model to assess declines in cognitive function as a function of age and to examine whether this decline is associated with pollution exposures. Specifically, we will consider models of the form:
\[ Y_{ij} = f_j(age_{ij}) + u_i + e_{ij}, \]  

(10)

where \( Y_{ij} \) is the MMSE outcome for subject \( i \) on day \( j \), \( e_i = 0, 1, \text{ or } 2 \) is a categorical exposure indicator defined by the tertile of a given pollutant, and \( age_{ij} \) is the age of subject \( i \) on day \( j \). This model accommodates the well-known nonlinear trend of MMSE score as a function of age, owing to an age threshold beyond which subjects experience cognitive decline. Further, it allows this trend to vary by low, medium, and high levels of chronic exposure. The subject-specific random effects account for correlation among repeated assessments taken on the same subject. We will consider extending this model to allow the functional form of \( f \) to vary smoothly with a continuous pollutant concentration, by assuming that the coefficients in a basis representation for \( f \) vary with the continuous exposure.

4.2.2.4 Project 4 (Viva Birth Cohort): The primary outcomes include somatic growth from birth up to year 7 of life (Aim 1), cardiovascular outcomes (blood pressure, step test) (Aim 2), and cognitive function (Aim 3). Exposure hypotheses include assessment of the health effects of exposure pre-natally, during the first 6 months of life, and over the course of the lifetime. In what follows we outline our approach to estimate associations between pollution and each outcome. Analyses to accomplish the aim of evaluating sources of susceptibility, vulnerability and protection (Aim 4) will use the same class of models as those outlined below, but with additional interaction terms between an exposure metric and each potential modifier.

**Aim 1 – Growth:** We will use growth curve analysis methods to analyze the association between growth and individual pollutants, sources, and mixtures. These models are longitudinal data models that allow individuals to have subject-specific rates of growth. Let \( Y_{ij} \) denote a growth outcome (e.g., weight-for-length or BMI z-score, change in BMI, birth weight for gestational age, etc.), and let \( polli \) denote an exposure metric of interest (we outline the specific choices for this covariate in Section 2.2). To start we will use a quadratic growth curve model, but will explore the adequacy of this assumed functional form for growth. This baseline model is:

\[ Y_{ij} = (\beta_0 + u_i) + (\beta_1 + v_i)ag_{ij} + (\beta_2 + w_i)age_{ij}^2 + \beta_3 polli + \beta_4 polli * age_{ij} + \beta_5 polli * age_{ij}^2 + ... + e_{ij}. \]  

(11)

The \( u_i, v_i, \) and \( w_i \) are subject specific random effects, and we will evaluate whether all three of these terms are necessary for our data. The hypothesis of interest corresponds to a test of \( H_0 : \beta_3 = \beta_4 = \beta_5 = 0 \). If the quadratic model does not fit the observed growth trajectories well, we will also consider nonparametric growth curve models, which fall within the class of generalized additive mixed models (see Section 3.1.1 of this Core).

\[ Y_{ij} = f_{q_j} (age_{ij}) + u_i + ... + e_{ij}. \]  

(12)

where \( e_i = 0, 1, \text{ or } 2 \) is a categorical exposure indicator defined by the tertile of a given pollutant. This assumes different growth patterns for subjects having exposures in different tertiles of exposure, and interest focuses on whether \( H_0 = f_0 = f_1 = f_2 \), which can be tested using the difference of model deviances. We will also consider extensions of this model that allow the functional form of \( f \) to vary smoothly with a continuous exposure, which can be accomplished by assuming that the coefficients in a basis representation for \( f \) vary smoothly with the continuous exposure. This will avoid loss of power due to categorization of continuous exposure. Primary exposure hypotheses for growth will focus on pre-natal exposures, by trimester, and on exposure
in the first 6 months of life. As noted in 2.2, associations with PM$_{2.5}$, O$_3$ and BC will be estimated using both temporal and spatial variation in estimated trimester-specific and 6-month exposures. Associations with other individual pollutants, sources, and mixtures will be estimated using temporal variability in the Supersite measurements.

**Aim 2 – Cardiovascular Risk:** We will conduct longitudinal analyses of the association between blood pressure (both diastolic and systolic) at birth, 6 months, year 3, and year 7, and the trend in blood pressure from birth to 7 years. These analyses will use linear mixed models similar to those presented in the previous sub-section. Care will be taken to check the model assumptions for the form of BP trajectory, and we will consider the more flexible nonparametric growth curve models outlined above if necessary. Primary analyses will again focus on the pre-natal period and exposure during the first 6 months of life, and we will use the same exposure covariates as described above for somatic growth. We will also consider two additional potentially important time windows of exposure for this outcome. First, we will consider cumulative lifetime exposure from the time of birth up to the time of BP measurement. For this metric, the model takes the slightly different form:

$$Y_{ij} = (\beta_0 + u_i) + (\beta_1 + v_i)age_{ij} + (\beta_2 + w_i)age_{ij}^2 + \beta_3poll_{ij} + ... + \varepsilon_{ij}$$ (13)

where here $poll_{ij}$ denotes the cumulative lifetime exposure for subject $i$ at measurement $j$. This slightly different formulation reflects the fact that the exposure metric is not only subject-specific, but also time-varying. Second, we will consider the acute effects of short-term exposures. These assessments will use model form (3), but with short-term exposures (on the order of daily to weekly) calculated from either the Supersite, spatio-temporal models, or satellite data, depending on the pollutant (see Section 2.2). Finally, in order to investigate whether life-time exposures have an effect beyond that from exposures during early developmental periods, we will consider multi-exposure models that contain both pre-natal (or 6 month exposure) and life-time exposure.

**Aim 3 – Cognition:** We will first assess associations between exposures and cognition at 6 months and at 7 years of age using linear regression models with cognitive testing score as the outcome and an exposure metric and confounders as independent variables. For the 6 months outcomes, separate models will be run using estimated average exposure (individual pollutants or source contributions; see Section 2.2) for each trimester and for the first 6 months of life. We will assess effect modification of PM$_{2.5}$ by pollutant mixture, with pollutant mixture defined at the level of zip code. For the 7 year outcomes, single-exposure models will include those that use trimester-specific, first 6 months, and 7 year lifetime exposures. We will also consider multi-exposure models that contain both an early (trimester-specific or first 6 months) and lifetime exposure.

**4.2.2.5 Project 5 (The National Study):** Because this Project is also inherently statistical, complete details of the proposed models are outlined therein. Here we briefly outline the chosen exposure metrics to be used for exposure hypotheses related to pollutant sources and mixtures. For hypotheses focusing on estimation of the health effects associated with short-term exposure to pollution sources, we will select a subset of geographical locations (e.g., counties). Separately, for each geographical location, source apportionment analyses will be performed on the daily pollutant concentrations. Subsequently, the estimated daily source contributions will be used in location-specific semiparametric time series models for mortality and hospital admissions. Results will be interpreted for each location individually, with attention focusing on whether one
or more sources are consistently associated with mortality/hospitalization across locations. Exposure hypotheses will also focus on the chronic effects of long-term exposures to pollutant mixtures, by applying clustering algorithms (Section 4.4.2) to location-specific long-term averages of pollution components. This analysis will partition study locations into groups of areas having similar long-term exposure profiles. The second stage of analysis will fit regression models estimating the effects of long-term exposure to PM$_{2.5}$ mass. We will then investigate whether these effects vary across groups of locations having the same mixture profile (see Project 5 for details).

4.3. Additional Analytical Considerations

4.3.1 Methods for Missing Data: There will likely be two different types of incomplete data/missingness. In some limited cases, there will be missing pollution data. This missingness is usually completely random and does not bias the results. There can also be dropout in the study, whereby subjects will come to the first several visits but not the later ones. Here, the correct strategy for dealing with dropout will depend on the missing data mechanism. If the data are missing at random (i.e. dropout is associated with other observed variables) or missing completely at random (dropout has no relation to the data), then the mixed hierarchical models we have proposed remain valid and will give unbiased estimates of effects when applied to all available data. In the conditional likelihood approaches, such as case-crossover settings, we can use the methods of Rathouz$^{42,43}$ to obtain valid estimates of the associations of interest. Further, we will systematically ascertain the reasons (including health reasons) for a participant missing a visit or missing a measurement. If Project investigators feel strongly that dropout is associated with the missing data (i.e., a person does not come to their visit due to their health at the present time), then we will conduct a sensitivity analysis that uses the hierarchical model as the baseline analysis, and then builds in models for the missing data mechanism. For instance, we can apply the shared parameter approach of Wu and Carroll$^{44}$ relatively easily within the hierarchical mixed model approach, which can be programmed in SAS PROC NLMIXED.

4.3.2 Methods for Multiple Comparisons: Multiple comparison issues may arise in testing for associations between multiple outcomes, multiple pollutant exposures, and multiple pollutant lags. We deal with these multiple comparisons in several complementary ways. First, when we are considering the effects on many different endpoints simultaneously, we will consider recently developed statistical procedures for controlling the false discovery rate (FDR) for multiple outcomes.$^{45}$ This approach avoids the overly conservative approach of Bonferroni and other traditional procedures. Second, we will use model-based approaches such as factor analysis/structural equations models (SEMs) for multiple outcomes or multiple pollutants and distributed lag models for multiple-lagged exposures that reduce the dimension of the problem and hence reduce the number of comparisons. Another novel opportunity for multiple comparison control in this PM Center is the use of one Cohort to serve as the replication sample for another. This is particularly relevant in the NAS and Framingham Cohorts, as there will be multiple outcomes that are common to both studies.

4.4. Methodological Research: With Center investigators, we identified several areas requiring methodological advances for the fullest use of the Center data during the next 5 years.
4.4.1. Spatio-temporal Modeling

Integrating Spatial-Temporal Exposure Information across Multiple Data Sources: As part of our ongoing and previous Centers we have pursued multiple approaches to refine location- and time-specific estimates of PM exposures based on use of different data sources in parallel. First, in order to cover the study Region proposed by the three Cohort studies (NAS, Framingham and Viva), we have extended the spatial coverage of our model by incorporating additional monitoring data on BC levels. This work extended the coverage of our existing data from the greater Boston area out to northern Connecticut to the South and to southern New Hampshire and Maine to the North. Second, our existing BC spatio-temporal model is calibrated to 24-hr monitoring data from more than 120 locations. We have also collected 7-day integrated indoor samples for a subset of NAS study subjects. In the Exposure Core we proposed to use satellite AOD data to obtain spatially- and temporally-resolved PM concentrations in the Region. One can view these different data sources as layers of information on spatio-temporal variation in exposure, each having its own set of strengths and weaknesses. As a major initiative of the Biostatistics Core, we propose to develop methods and models that appropriately integrate information on exposure from different data sources, and also appropriate computational strategies necessary to analyze the resulting large exposure datasets. For example, we have preliminary results for a method for pooling 24-hr monitoring data used in our original BC models with the 7-day integrated indoor monitoring BC data from the NAS monitors. Incorporating the integrated data nearly doubles the number of unique spatial locations at which data are available. In order to pool data from 24-hr and 7-day temporal scales, we specify a model for the 24-hr values, and then derive the likelihood contribution from an integrated 7-day measure by aggregating the original model over the 7-day period. This is essentially the strategy used to model change-of-support spatial data, in which the likelihood contributions for area-based concentrations are derived from an originating model for geo-referenced point data.46

Spatio-temporal Model for Ozone: We propose to develop an O3 model analogous to the BC model to predict exposures throughout the Region. Because there are not enough sites within the urban areas where many of the participants live, we will use O3 monthly measurements from the 40 proposed HSPH multi-pollutant sites to further develop and characterize the model. The model will be a land-use regression model that combines effects of traffic-related GIS-variables and smoothly varying spatial terms in an additive model framework. We are optimistic that our approach will yield adequate fits to the data. However if the model fits the O3 monitoring data significantly worse than our existing BC model fits the BC data (as quantified by out-of-sample validation), we will fall back on two alternative strategies, which are to: 1) investigate the performance of simpler kriging models; and 2) rely more heavily on the long-term spatial models for O3 described in Section 4.4.3

4.4.2 Clustering Complex Mixtures: A major goal of the proposed Center is to identify pollutant mixtures associated with health effects. We will perform two types of clustering: 1) for acute effects, we will group days based upon their multi-pollutant profiles using temporal clustering; and 2) for chronic effects, we will group geographical areas according to their yearly average pollutant concentration profiles. For the Cohort studies (NAS, Framingham and Viva) and the National Study, we will apply such spatial clustering at the zip-code level. We will then investigate whether the health effects associated with short or long-term exposure to PM is modified by the pollutant mixture-based grouping. We have collaborated closely with members
of the Center's Exposure Core to apply K-means clustering algorithms, implemented in the SAS program PROC FASTCLUS, to daily exposure data from the Boston Supersite (see Exposure Core). We will continue to work closely with the Exposure Core to refine methods for clustering pollution mixtures, both between days and across locations. In addition to K-mean clustering, we will also consider model-based and hierarchical clustering as viable methods to effectively identify distinct pollution events. Given a set of observations where each observation is a \(d\)-dimensional real vector, K-means clustering partitions the \(n\) observations into \(k\) sets (\(k < n\)), so as to minimize the within-cluster sum of squares (WCSS),

\[
\arg\min_{\mu} \sum_{i=1}^{n} \sum_{j=1}^{k} ||x_{ij} - \mu_i||^2,
\]

where \(\mu_i\) is the cluster-specific mean vector. An advantage of K-means clustering is its simplicity, rendering it fast enough to apply to large datasets. Disadvantages include the need to specify the number of clusters, and the fact that it yields clusters having simple spherical shapes in multivariate space. Accordingly, we will also consider model-based clustering, which allows data-driven choice of the number of clusters using the Bayesian Information Criterion implemented via the mclust package in R. We will compare the performance of K-means clustering to other existing methods with respect to maximizing between-cluster variability and minimizing within-cluster variability, for exposure concentrations and component ratios that serve as indicators of pollution events (see Exposure Core). We will also simulate exposure data from a known set of clusters, and evaluate each method's ability to retain the correct number of clusters.

### 4.4.3 Long-term Spatial Models:

We will develop long-term multi-pollutant spatial models for data generated by the multi-pollutant spatial monitoring effort conducted in the proposed Center (see Exposure Core for details). Briefly, in each of the 40 sites, 4 monthly samples will be collected per year in January, April, July and October for four years. Because research has shown that using a kriged or otherwise interpolated estimate of exposure at a given location generally outperforms a nearest-monitor strategy,\(^{47,48}\) we will develop models that smooth long-term pollutant concentration levels over space. We will assume a different spatial surface for each of the four seasons. For some health analyses, we will extrapolate estimated spatial estimates of pollutant levels to the 1999-2009 period. Therefore, we will calibrate our long-term spatial estimates by the central site readings on a given day. For a given pollutant and a given season, our starting point will be models of the form:

\[
Y_{st} = g(s) + \beta_{central_t} + \epsilon_{sj} \tag{14}
\]

where \(Y_{st}\) is the pollutant concentration at spatial location \(s\) at time (year) \(t\), \(g(s)\) is the smooth spatial term, and \(\beta_{central_t}\) is the long-term average for that pollutant over the time during which \(Y_{st}\) was collected. This simple approach calibrates predictions for earlier years in the presence of a decreasing trend in pollutant levels. After applying this approach to data on each pollutant separately, we will explore the possibility of extending this model to a multi-pollutant one that predicts the multivariate exposure in a unified framework. This will involve specifying correlations among the coefficients used in the basis representations of \(g(s)\) for the multiple pollutants.

### 5. EXPECTED RESULTS

The statistical methods, computationally efficient algorithms, and software packages developed by our Biostatistics group have been critical to our current Center's ability to address key
scientific issues pertaining to PM health effects. These include studies of measurement error effects and of links between effects and specific sources, and issues related to estimated exposures in health effects analyses. Much of this work has been included in recent EPA PM Criteria Documents and has been critical to the validation of our epidemiological studies. This Core will continue to develop and apply state-of-the-art analytical tools to our epidemiological, toxicological and exposure data. We will also continue to develop new statistical approaches to investigate the health effects of air pollution. As in the past, Core investigators will disseminate new methods in the statistical literature, and the Center will share these new tools with other researchers in the PM field including other Centers.

6. GENERAL PROJECT INFORMATION

The Core will be led by Dr. Brent Coull, who will be assisted by Core Co-Leaders Dr. Francesca Dominici (PI, National Study Project) and Dr. Joel Schwartz (PI, NAS Project), and Co-Investigator Dr. Antonella Zanobetti. Drs. Coull, Schwartz, and Zanobetti have worked closely together for over 10 years and have co-authored many papers. In addition, they have worked closely with the epidemiologists, exposure assessors and toxicologists of the existing Center and these collaborations have produced many interdisciplinary peer reviewed publications. Dr. Dominici has collaborated with Drs. Schwartz and Zanobetti on several projects over the past 10 years, and has co-organized several scientific workshops with Dr. Coull. In addition, Drs. Dominici and Coull are the co-Directors of the Environmental Statistics Program at the Harvard School of Public Health. Detailed descriptions of each investigator’s role in the Core are provided in the budget justification. Meetings with each Project in the Center will be held on a bi-weekly basis, and Core members will report on progress made by the Core to all Center investigators at the Center-wide meetings (see Administrative Core). This Core will provide study design and data analysis support for all five projects of the proposed Center. Core members will assist project investigators in the design of the proposed studies and will participate in data analysis. To the extent allowable by design and outcome commonalities among the five Projects, Core investigators will ensure that a unified approach to data analysis, in terms of modeling strategy and choice of data transformations, is applied to all Center data.

REFERENCES


44. Wu MC, Carroll RJ. Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics* 1988; 44:175-188.

