

## **ABSTRACT**

The Biostatistics Core will provide centralized statistical and analytical expertise to all Center projects. The 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 bring 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, 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 substantial design consultation and analytical work, and; (2) training for investigators in both statistical issues involved in the data analysis as well as in SAS and Splus/R software. In addition, the Core will (3) conduct mission-related methodological research when existing methodology does not fully address the scientific question of interest.

A critical component in all of the projects is power and sample size calculations. Prospective calculations allow project investigators to be confident that the Center projects will have high power to detect meaningful differences. Core investigators 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% or 90%.

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. Data analyses will apply good exploratory data analysis techniques, such as univariate explorations of the data, distributional checks, and outlier identification to data from all projects. For model building, residual analysis and other model diagnostics to confirm model fit, identify possible nonlinear relationships between predictors and outcomes, and identify highly influential data points will be routinely employed as part of a sound data analysis strategy. Once the data have been checked and modeling assumptions verified, primary analysis methods will include ANOVA and regression techniques, with the particular form and correlation structure of the data dictating the particular method. The main methods of analysis will be linear models/generalized linear models, multi-way ANOVA, semi-parametric regression modeling, and mixed/multivariate models for correlated responses.

Methods developed by Core investigators with support from the current Center will play a large role in future Center investigations. These techniques include smoothing methods, distributed lag models, exposure measurement error corrections, and case-crossover analyses. Future methodological developments will focus on spatial modeling of pollution, methods to address model uncertainty, and methods to assess the health effects of complex pollution mixtures.

## 1. OBJECTIVES

The Biostatistics Core will provide centralized statistical and analytical expertise to all projects. The 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 bring 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, meta-analysis, structural equation models, and Bayesian data analysis. The Biostatistics Core will provide:

- Support for statistical analysis for all five proposed projects, including substantial design consultation and analytical work;
- Training for graduate students and investigators in both statistical issues involved in the data analysis as well as in SAS and Splus/R software, and;
- Mission-related methodological research when existing methodology does not fully address the scientific question of interest.

## 2. INTRODUCTION

Over the last five years, current Center investigators Brent Coull, Joel Schwartz, and Antonella Zanobetti have provided statistical support and addressed methodological issues arising from challenging data analyses. As a result of these activities, a large spectrum of data analysis approaches was applied and several new methodologies were developed. Because these approaches and new tools will be employed by the proposed Center projects, a brief description of these analytical techniques is presented below.

## 3. PROGRESS

**3.1. Data Analysis:** Because an important goal of the existing Center has been the investigation of associations between biologic responses and exposures, the main thrust of data analysis has been the use of regression techniques and extensions thereof. However, because the nature of the outcomes, study design, and exposure characterization vary by study, analytical approaches differ by project.

**3.1.1. Exposure metrics of increasing sensitivity:** Both the ongoing and the proposed animal and human toxicological experiments use a partial factorial design. In these studies, subjects are randomized to receive either control (particle-free air) or concentrated air particles (CAPs) exposures, either within a straightforward factorial structure, or a more complicated cross-over design. On a given exposure day, however, the Harvard Ambient Particulate Concentrator (HAPC) produces essentially random CAPs exposures, both in terms of the particle mass levels and composition. This presents statistical challenges above those encountered in classic toxicology experiments of controlled exposures. To analyze these data, we have conducted multiple analyses that use exposure metrics of increasing sensitivity. First, we begin by using (potentially multi-way) ANOVA techniques that treat exposure as a treatment. That is, we assess overall differences between CAPs and filtered air responses (i.e. a binary exposure covariate).

Second, to assess univariate associations between CAPs mass levels or composition and health, we conduct single-pollutant analyses in which a separate regression model is fitted using 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 conduct multiple pollutant analyses to investigate joint effects of multiple pollution sources. Toward this end we have linked biological outcomes to predictors such as: (1) multiple tracer elements of previously defined pollution sources<sup>1,2,3</sup>, or (2) factor scores obtained from source apportionment factor analysis to represent multiple pollution sources in ambient exposures<sup>4,5</sup>.

As noted above, these random exposures are typically nested within a complex experimental design, such as a cross-over or repeated measures design. Thus, although we refer to these three levels of analysis as “ANOVA” and “regression”, this strategy of increasing the sensitivity of the exposure metric is usually nested within regression extensions that respect the design of the study and the correlation structure of the data. We briefly review the more commonly-used extensions in Sections 3.1.2-3.1.6, which apply more generally across all areas of the Center including: epidemiology, exposure assessment, and toxicology.

**3.1.2. Multilevel Random Effects Models for Longitudinal and Clustered Data:** Throughout the Center, data are often collected longitudinally within a subject or otherwise clustered (e.g. by city). For example, in the exposure studies, both personal and ambient concentrations are collected repeatedly over time<sup>6,7,8</sup>. In the animal studies, biologic outcomes are measured repeatedly on the same animals under multiple exposures<sup>3,9</sup>, and in epidemiological studies, researchers use panel study designs<sup>10,11</sup> or data from multiple cities<sup>12,13,14,15</sup>. In such cases, it is important to account for the dependence among measurements taken on the same cluster. We have used hierarchical models with random effects to account for repeated measures within a longitudinal design and other forms of clustering. These regressions combine fixed effects, which are the predictors in traditional regression models, with random effects that account for correlation among observations from the same cluster<sup>16,17</sup>. Specifically, let  $Y_{ij}$  be the response from subject  $i$  on day  $j$ . For concreteness, we consider the epidemiological bus study setting and consider covariates  $pollution_{ij}$ ,  $gender_i$ ,  $age_i$ , and  $temperature_j$ , which include both time varying (pollution, temperature) and between-subject (age, gender) covariates. The models apply for repeated measures in the toxicological studies as well. We have relied heavily on models of the form:

$$Y_{ij} = (\beta_0 + u_i) + \beta_1 gender_i + \beta_2 age_i + \beta_3 temp_j + \dots + (\beta_p + v_i) pollution_j + \dots + \varepsilon_{ij} \quad (1)$$

Here  $u_i$  is a subject-specific intercept that reflects unexplained heterogeneity in subjects’ overall level of outcome, and  $v_i$  is a subject-specific slope that reflects heterogeneity in susceptibility to pollution exposure. We make the standard assumptions that  $(u_i, v_i)$  are generated from a multivariate normal distribution with general variance covariance matrix, although we use newly developed diagnostic methods to assess the validity of this assumption (See Section 3.2.5). Such models for continuous outcomes can be easily fitted using standard software, such as PROC MIXED in SAS and the lme() function in S-plus/R. Extensions to more than two levels of the hierarchy (i.e. repeated measures within subject within city) are handled similarly, with the model containing both subject-specific and city-specific random effects. For discrete responses, such as dichotomous outcomes or counts, we use generalized linear mixed models (GLMM) as

in Gold et al.<sup>11</sup> and use PROC NLMIXED in SAS to fit these models. This procedure has the advantage over other GLMM macros that estimation is based on numerical integration of the likelihood, which avoids bias associated with rougher approximations of the likelihood employed by other procedures (i.e. SAS macro GLIMMIX). We have also used generalized estimating equations (GEEs) for fitting longitudinal discrete responses<sup>9</sup>.

**3.1.3. Susceptibility:** We are also interested in whether effect modification by characteristics of the subject, such as diabetes status and medication use, among others, accounts for some of the unexplained heterogeneity in either the overall outcome level ( $u_i$ ) or pollution effect ( $v_i$ )<sup>11</sup>. This question is easily incorporated in the hierarchical framework by adding main effects for those characteristics and interactions with pollution, respectively. For instance, in model (1), we address this by specifying that the mean of the random slope terms varies according to this subject characteristic:  $(\beta_p + \beta_{p+1} \text{modifier}_i + v_i) * \text{pollution}_j$ . These terms amount to pollution\*modifier terms in the mixed models.

**3.1.4. Generalized Additive Models:** In many cases, the assumption of linearity between the response and a covariate in a linear regression model is overly restrictive and may bias estimates of exposure effects. Examples of how this assumption could be violated in Center settings include complex confounding effects of temperature or season in epidemiological studies<sup>18</sup> and non-linear profiles of animal cardiac and pulmonary parameters over an exposure period in the animal studies<sup>19,20</sup>. If the functional forms of these relationships could be reliably determined, one could reflect this in the model using variable transformation. However, these associations are often complex nonlinear functions, and more flexible methods are necessary. An alternative tool that accounts for this complexity is the Generalized Additive Model (GAM)<sup>21</sup>. This powerful modeling approach has become very popular in recent years due to the flexibility of these models in capturing non-linear relationships between variables. Suppose  $Y_i$ ,  $i=1, \dots, n$ , represent  $n$  independent observations, and let  $X_{i1}, \dots, X_{ip}$  denote a corresponding set of  $p$  covariates of interest. These covariates may represent one or more exposure measures as well as potential confounders such as time or co-pollutants. A linear regression model for  $Y_i$  is:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip} + \varepsilon_i \quad (2)$$

where  $\beta_0$  is the overall intercept and  $(\beta_1, \dots, \beta_p)$  represent the linear effects of the  $p$  covariates on the response  $Y$ . The idea is to replace linear effects  $\beta_j X_{ij}$  in model (2) with an arbitrary smooth function  $f(X_{ij})$ , which is then estimated from the data. We have used semiparametric regression, which are GAMs that contain both parametric terms  $\beta_j X_{ij}$  and smooth functions  $f(X_{ij})$ , extensively in daily mortality studies<sup>18</sup> as well as for analyzing cardiac HRV and pulmonary data from the animal studies<sup>19</sup>. Dominici et al.<sup>22</sup> noted that the most popular program for fitting GAMs, Splines, reports incorrect results when more than one smooth term is included in the model and the two curves are collinear. We have spent a great deal of time assessing the implications of this discovery for results generated by the current Center. Specifically, we re-analyzed our 10 city mortality study, the Six City time series study, the Six City Source Apportionment Study, our Hospital Admissions studies, and the long term distributed lag models from the APHEA study using different convergence criteria and natural splines. In addition, we have developed software that uses penalized splines<sup>23</sup>, an alternative approach that avoids the problems incurred by Splines. We defer detailed discussion of this work until Section 3.2.2.

**3.1.5. Time to Event Data:** Events, such as cardiovascular deaths, are treated as censored survival time data using standard Cox regression models. With this approach, all individuals can be included in the cohort, making complete use of the data, and their actual times of the events (MI, death, etc.) can be analyzed in relation to risk factors measured at baseline (standard cardiovascular risk factors, baseline measurements of intermediate endpoints, and air pollution). To allow for the clustering of observations by community, we use the random effects Cox model<sup>24</sup>, recently developed for the reanalysis of the Six Cities and ACS cohort studies of the chronic effects of air pollution on mortality.

**3.1.6. Meta-Analyses of Multiple Studies:** Some Center analyses<sup>12,13,14,25</sup> pool data from several locations or several existing studies to assess the weight of evidence across multiple settings. We have used the hierarchical two-stage approach of Berkey et al.<sup>26</sup> for this purpose. In the first stage, we fit identical models for the outcome measures from the different locations/studies. These city-specific models can be complex, possibly including random effects to account for correlated data or spline terms for non-linear effects. In the second stage, we pool the slopes from the different studies to examine effect modification. That is, do city-specific particle effects vary systematically according to measured characteristics of the city, such as region of the country. Specifically, the second stage model is:

$$\widehat{\beta}_i = \alpha_0 + \alpha_1 * w_i + \delta_i + \varepsilon_i, \quad (3)$$

where  $\widehat{\beta}_i$  is the estimated PM slope in study  $i$ ,  $\alpha_0$  is the estimated effect of PM in the absence of the effect modifier  $w_i$ ,  $\alpha_1$  is the change in effect of PM in the presence of the effect modifier,  $\delta_i$  is a random error term with known variance equal to the estimated variance of the parameter estimates obtained from the first stage of the model, and  $\varepsilon_i$  is a random error term with unknown but estimable variance. This meta-analytic framework is a flexible way to summarize the regression coefficients for each outcome. If the effect modifier term is left out of the model, the meta-analysis amounts to calculating an overall estimate that pools information from all of the studies.

**3.2. Methodological Development:** As part of the current Center (in addition to providing study design and data analysis support) we have developed new statistical methodology for situations for which existing methods inadequately address the scientific question of interest. These developments are summarized below.

**3.2.1. Measurement Error:** In addition to the current Center's individual exposure assessment studies that address the extent of measurement error in PM epidemiology<sup>6,7,27</sup>, we have focused on measurement error from a methodological perspective as well. In doing so, we have developed new methods to correct for measurement error in hierarchical models<sup>14</sup>. We showed that existing standard two-stage estimators will be biased in the presence of exposure measurement error, and that this bias can be away from the null hypothesis of no effect. We proposed two alternative methods, termed an intercept estimate and a slope estimate, for estimating the independent effects of two predictors in a hierarchical model. We showed both analytically and via simulation that the slope estimate gives essentially unbiased estimates even in the presence of measurement error, at the price of a moderate reduction in power. We applied

the new methodology to show that the estimated effect of fine particles on daily deaths, independent of coarse particles, was biased downward by measurement error in an original analysis that did not correct for measurement error. In a second application of the model, we used published data on the association between airborne particles and daily deaths in 10 US cities to estimate the effect of gaseous air pollutants on daily deaths. The resulting effect size estimates were very small and the confidence intervals included zero. We have also applied this approach to a reanalysis of the NMMAPS mortality study conducted by researchers at Johns Hopkins University. A paper reporting this NMMAPS re-analysis is now under review<sup>28</sup>. Finally, we have conducted a simulation study (manuscript in progress) to assess the amount and pattern of bias in health effect estimates if the personal-ambient relationships followed those reported in several individual exposure studies conducted in Boston and Baltimore as part of the current Center<sup>7,27,29</sup>. Results provide evidence that the gaseous pollutants are unlikely confounders of PM health risk estimates for these locations.

**3.2.2. Generalized Additive Mixed Models:** As noted in Section 3.1.2, mixed models represent an important tool for determining whether individuals with certain characteristics are more susceptible to the effects of airborne particles. However, classic mixed regression models are linear models, whereas we know that the effects of season and weather on health are often nonlinear. To enhance our ability to assess sensitivity while maintaining good covariate control, we have developed additive mixed models, which combine the hierarchical structure of model (1) with the nonparametric regression techniques of Section 3.1.4<sup>30</sup>. Specifically, consider covariate  $X_i$ , and let  $\kappa_k, k=1, \dots, K$ , be a set of  $K$  distinct values within the range of  $X$ . For a smooth function  $f(X_i)$ , the mixed model formulation of a penalized spline model for the curve specifies a piecewise polynomial fit:

$$f(x_i) = \beta_1 x_1 + \dots + \beta_p x^p + \sum_{k=1}^K u_k (x_i - \kappa_k)_+^p, \quad (4)$$

where,  $(x_i - \kappa_k)_+$  is defined as

$$\begin{cases} x_i - \kappa_k, & x_i > \kappa_k \\ 0, & \text{otherwise} \end{cases}$$

and the truncated polynomial coefficients  $u_k$  are independent normally distributed random effects with unknown variance  $\sigma_u^2$ <sup>31</sup>. As a result, we can simultaneously estimate nonlinear effects in the presence of a multilevel structure by specifying a single mixed model with both cluster-specific and smoothing random effects. We have recently applied this approach to estimate the association between  $PM_{10}$  and mortality. We have also conducted methodological research on effective computational strategies for fitting semiparametric mixed models. We have been part of the development team of *Semipar*, an Splus/R module for fitting a wide variety of semiparametric regression models based on penalized splines<sup>32</sup>, and have conducted a systematic investigation of Bayesian approaches to fitting generalized additive mixed models<sup>33</sup>.

**3.2.3. Mortality Displacement (Harvesting):** We have developed new methodology for assessing the “mortality-displacement”, or “harvesting” effect, of particulate pollution, which corresponds to the hypothesis that observed associations with mortality are due solely to the deaths of frail individuals. The *smoothed distributed lag model*<sup>34</sup> estimates the net effect of air pollution on daily deaths, which includes short-term rebounds due to mortality displacement as well as any longer-term lagged effect. Let  $\beta_l$  be the effect of PM at time lag  $l, l=0, 1, \dots, q$ . The

model specifies a smooth structure for  $\beta_l$  by placing a penalized spline structure on these coefficients:

$$\beta_l = \sum_{j=0}^p \tau_j l^j + \sum_{k=1}^K \nu_k (l - \kappa_k)_+^p \quad (5)$$

In addition, the model allows for non-linear effects of other confounding variables, such as temperature. Let  $z_t$  denote a set of confounder variables modeled linearly, let  $s_{jt}$  denote a set of variables to be modeled with a smooth function, and let  $x_t$  denote pollution concentration, all collected at time  $t$ . The full generalized additive distributed lag model is

$$g[E(y_t)] = \alpha + \gamma' z_t + \sum_{j=1}^d f_j(s_{jt}) + \sum_{l=0}^q \beta_l x_{t-l}, t = q+1, \dots, T, \quad (6)$$

where  $f_j$  is also estimated using penalized splines. We applied the model to data on the relationship between pollution and daily deaths in Milan, Italy<sup>34</sup>. This paper confirmed that, far from reducing effects, “harvesting resistant” estimates are higher by a factor of two. More recently, we used the distributed lag approach to examine the potential harvesting effect in 10 European cities. The findings from this study and others conducted within the current Center also provided no evidence to support a harvesting effect<sup>35,36,37,38</sup>.

**3.2.4. Case–Crossover Analyses:** The case-crossover design, introduced by Maclure<sup>39</sup>, is a method for investigating the acute effects of an exposure. In the case-crossover approach, a case–control study is conducted in which each person who had an event is matched with herself/himself at a nearby time period during which she/he did not have the event. The subject’s characteristics and exposures at the time of the case event are compared with control periods in which the event did not occur. Each risk set consists of one individual as that individual crosses over between different exposure levels in the case and control time periods. These matched pairs may be analyzed using conditional logistic regression. Multiple control periods may be used. In recent years, this approach has been applied to the analysis of the acute effects of environmental exposures, especially air pollution. Applied to the association of air pollution with risk of death, the approach presents several advantages. Because each subject serves as her/his own control, the use of a nearby day as the control period means that all covariates that change slowly over time, such as smoking history, age, body mass index, usual diet, and diabetes are controlled for by matching. Therefore, the case-crossover design controls for seasonal variation, time trends, and slowly time varying confounders by design, because the case and control periods in each risk set are separated by a relatively small interval of time. In particular, season and time trends are controlled by matching. Bateson and Schwartz<sup>40,41</sup> demonstrated that, by choosing control days close to event days, even very strong confounding of exposure by seasonal patterns could be controlled by design in the case control approach. While it is straightforward to sample control days in a manner that removes seasonal confounding, there can be a subtle selection bias in these analyses. Several approaches have been shown to address this problem, including the time stratified approach of Lumley and Levy<sup>42</sup>. We have recently applied this approach to the issue of confounding by gaseous co-pollutants, in an analysis of 14 cities where control days were matched on the level of the co-pollutant. Significant associations with PM<sub>10</sub> were seen in all analyses<sup>43</sup>.

**3.2.5. Diagnostics for Multilevel Random Effects Models:** An important assumption in the use of linear mixed effects models is the normality of the random effects and the errors. For

instance, in random slopes models that assess population heterogeneity in susceptibility to pollution exposure (a National Research Council defined critical need), the distributional assumption of normality for susceptibility serves to pool information on the pollution slopes across subjects and shrink individual estimates toward the overall mean susceptibility. However, if the distribution of susceptibility is skewed such that a small subset of subjects are highly susceptible to exposure, the normality assumption results in excessive shrinkage of large estimated slopes toward the overall mean<sup>44,45</sup>, leading to underestimated effects for these subjects. With support from the current Center, we developed diagnostic graphics and hypothesis tests to assess the validity of distributional assumptions in linear mixed models<sup>46,47</sup>. Houseman, Coull, and Ryan<sup>46</sup> used these tests to conclude that three subjects in a study of the effects of PM<sub>10</sub> on pulmonary function in Utah Valley schoolchildren exhibited significantly higher susceptibility than the rest of the cohort.

## 4. RESEARCH METHODS

**4.1. Study Design:** Core members have worked closely with Center investigators to develop the study design for each proposed project. A critical component in all of the projects is power and sample size calculations. Prospective calculations allow project investigators to be confident that the Center projects will have high power to detect meaningful differences. 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% or 90%. Specific power and sample size calculations for the proposed experiments can be found in the individual project proposals. In general, we take the following approaches. For standard models such as ANOVA and linear regression, we use standard variance formulas for estimates to compute sample sizes and power for effect sizes observed in previous studies and pilot data. Power calculations for hierarchical models use variance formulas for linear regression and the fact that the mixed model can be expressed as a series of such regressions<sup>16</sup>. Thus, for a given effect size and assumed values for the within-subject and between-subject residual variances, we can carry forward the uncertainty of the estimated coefficients at each level of the model to obtain overall variances of the effect estimates of interest. For events, we use the equivalence between survival models and Poisson regression<sup>48</sup>, and use a simulation-based approach to assess power for a multilevel Poisson model. 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 investigators in this Center. We also use existing data to provide estimates of the amount of variability of exposure. In general we consider only the temporal component of exposure variability. For pollutants with large spatial components of variability, these power estimates would be conservative for the spatio-temporal methods outlined in Section 4.3.2. All power calculations are based on two-sided tests at the 0.05 level.

**4.2. Data Analysis:** We will apply good exploratory data analysis techniques, such as univariate explorations of the data, distributional checks, and outlier identification to data from all projects. For model building, we will employ residual analysis and other model diagnostics 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, we will apply ANOVA and regression techniques to Center data, with the particular form and correlation structure of the data dictating the particular method. The main

methods of analysis will continue to be linear models/generalized linear models, multi-way ANOVA, semi-parametric modeling, and mixed/multivariate models for correlated responses (For a full description of nonstandard techniques, see Sections 3.1.1-3.1.6). In addition, we also anticipate that statistical tools developed as part of our mission-related statistical research (see Section 4.3) will become staples of our toolbox for the analysis of Center data. We now give more detail for data analyses in each project.

**4.2.1. NAS Study (Project 1):** Analyses of the NAS Project 1 data will use linear regression, semi-parametric regression, Cox proportional hazard models for analyzing events and, for outcomes measured during two rounds of the NAS study, longitudinal models. Responses of interest will include heart rate variability (HRV), blood pressure, fibrinogen, measures of inflammation (CRP, ICAM, and VCAM), lung function, and events such as heart attacks. We will investigate genetic susceptibility in the above models by including interaction terms for pollution and an indicator of the GSMT1 or HO-1 polymorphisms in the models, as outlined in Section 3.1.3.

Because subject residences will be geocoded, we will also use spatial regression models to account for potential spatial correlation among outcomes. We can build spatial relationships into the proposed statistical framework in two different ways. First, we can assume the residual errors exhibit spatial correlation, with this correlation being a function of the distance between two observations. Such analyses of continuous outcomes can be implemented using the `sp(exp)` syntax in PROC MIXED. Second, we can accommodate any potential residual spatial patterning of an outcome within the Boston area using geospatial models<sup>49</sup>, which specify that the mean response depends on exposure, other covariates, and a two-dimensional function of location. This approach is essentially a two-dimensional version of the nonparametric regression models discussed in subsection Section 3.1.4 of this Core. As in the one-dimensional case, we will use a penalized spline formulation for this surface, allowing us to incorporate the spatial structure in the mixed model framework<sup>50</sup>. Finally, we will also use spatial-temporal models for refining estimates of exposure for pollutants exhibiting large amounts of spatial variability, such as black carbon as a marker of traffic pollution. We defer a detailed description of this approach until Section 4.3.2.

**4.2.2. Bus Study (Project 2):** We will consider repeated measures (and as an extension, multilevel) mixed models for assessing pollution effects and effect modification on biologic outcomes of interest within a cross-over design. Outcomes of interest in the bus study will include exhaled NO, HRV, markers of systemic inflammation in the blood (CRP, ICAM, VCAM, IL-6), and blood pressure. For cases in which responses are recorded semi-continuously (such as five minute HRV measures) within an exposure period, with exposures repeated on each subject (such as in the proposed cross-over design), the data represent multiple levels of nesting with time nested within exposure, and exposure nested within animal. In such cases, we will employ multilevel versions of the hierarchical models described in Section 3.1.2. We will use standard model building procedures to arrive at a model that specifies an appropriate correlation structure among responses at each level of the hierarchy. For instance, in our previous analyses of cross-over concentrator data<sup>19,20</sup>, we specified a serial correlation structure for HRV observations recorded within the same six hour exposure in combination with animal effects to account for correlations among observations taken on the same animal but during different

exposure periods. Furthermore, the similarities between the animal and the bus cross-over study designs will allow us to use the same multi-tiered strategy that uses exposure metrics of increasing sensitivity in this project. That is, we will first use (potentially multi-way) ANOVA techniques that treat pollution exposure as a binary treatment (traffic particles plus gases versus gases only). Second, to assess how these differences vary according to the amount and chemical properties of traffic particles exposure, we will conduct univariate analyses in which a separate regression model is fitted using biologic response as the dependent variable and univariate pollution concentrations as the exposure metric. Third, to confirm univariate findings, we will conduct multiple regression analyses to investigate joint effects of multiple pollutants. One unique feature of the bus study will be the measurement of continuous biological outcomes and pollution concentrations, which makes it possible to examine the lag structure between exposure and response. Our previous work indicates that the temporal relation between cumulative exposure and cardiac outcome may be shorter for some outcomes (e.g. HRV) than for others (e.g. blood pressure). Lagged associations will be examined using polynomial distributed lag models<sup>34</sup>, which do not require multiple comparisons at different lags. Furthermore, continuous exposure assessments will enable us to address collinearity problems, since associations among pollutants are well-known to have strong diurnal patterns, with some associations reversing direction throughout the course of a day. We will use multi-pollutant generalized additive distributed lag models in these settings, in which the time-course of effects for each pollutant follows the distributed lag structure described in Section 3.2.3.

**4.2.3. Human Concentrator Studies (Project 3):** Statistical analysis for the human exposure studies will follow a multi-tiered strategy that successively uses exposure metrics of increasing sensitivity. The first stage will employ repeated measures ANOVA models containing a random effect for subject and a categorical variable for exposure group (filtered air, ultrafine, fine, coarse), to assess differences among groups. Standard multiple comparisons procedures, such as Dunnett's and Scheffe's procedures, will be employed to adjust for comparisons of multiple exposure groups. Second, to assess the effects of PM composition on measured outcomes, single pollutant dose-response analyses will be conducted in which a separate linear mixed regression model is fitted using biologic response as the dependent variable, subject as a random effect, and either mass, particle number, or a single elemental concentration as the exposure metric in the model. Third, to assess the relative contributions of multiple pollution sources on health, multiple pollutant regression analyses will be conducted using exposure metrics. For outcomes recorded semi-continuously within an exposure, hierarchical linear models will be developed to account for the multiple levels (time within exposure within subject) of data. Such techniques allow assessment of differences among within-exposure slopes across exposure groups.

**4.2.4. Animal Concentrator Studies (Project 4):** We will continue to apply ANOVA and regression techniques to data generated from the animal concentrator studies, with the particular form and correlation structure of the data dictating the use of a particular method (See Section 3 for a full description). Such models will take advantage of the partial factorial experimental designs for the concentrator studies. The main methods of analysis will continue to be linear models, generalized linear models, multi-way ANOVA, semi-parametric models, and mixed/multivariate models for correlated outcomes. Responses of interest will include hypertension, blood pressure, BUXCO outcomes, chemiluminescence, and BAL outcomes. All of these models can now be fit in standard software, such as PROC MIXED in SAS or Splus/R. As

in Projects 2 and 3, we will continue to use the multi-tiered strategy that successively incorporates more refined measures of exposure to assess the relative contribution of different sources to observed effects

**4.2.5. TERESA Study (Project 5):** Project 5 will involve animal exposure studies in which rats are randomly divided into treatment groups, with all rats from a given group receiving a given exposure scenario (i.e. filtered air sham, primary gas and particle emissions, primary plus secondary particles—with and without neutralization by ammonia, secondary without primary particles) of a given pollution source (traffic emissions). In addition, this protocol will be replicated for different strains of rats (healthy and spontaneously hypertensive rats). Outcomes of interest on each animal will include chemiluminescence, bronchoalveolar lavage (BAL), BUXCO pulmonary measures, blood parameters, and EKG recordings. We will use multi-way ANOVA to assess differences among groups defined by type and source of pollution, as well as rat type. We will use PROC GLM in SAS to first assess overall differences among levels for each factor, and then use multiple comparisons procedures, such as Scheffe's multiple comparison procedure for cohort groups or Dunnett's procedure when interest focuses on making comparisons against the filtered air control, to assess differences between pairs of factor levels and to examine which treatments/rat cohorts are significantly different from others for a given outcome. Although there is a target dose of pollution for each exposure, there will be some variability in these levels across experimental runs. If there is significant dose variability, we will also extend the multi-way ANOVA analyses to the regression setting to assess dose-response slopes for each exposure treatment and rat group. Such analyses will enable researchers to assess the toxicity of each treatment combination in each cohort.

**4.3. Methodological Research:** Through our work with current Center investigators, we have identified several areas that will require methodological advances for the fullest use of the data from the proposed Center during the next five years.

**4.3.1. Structural Equations Models (SEMs):** We will investigate the use of Structural Equation Models (SEMs) to investigate relationships among multiple exposures and multiple health outcomes. SEMs are a class of covariance structure models that can be used to show path diagrams and to simultaneously model multiple surrogates of both exposure and outcome. Used extensively in social sciences and, more recently, in environmental epidemiology to investigate the health effects of lead<sup>51</sup> and methylmercury<sup>52</sup>, SEMs reduce the number of multiple comparisons made on multiple outcomes and adjust for covariate measurement error. The models are often represented as path diagrams, allowing one to fit models that specify outcomes as lying on a causal pathway between exposure and other outcomes. The models also allow for the specification of multiple surrogates of exposure on these pathways. In this framework, we will obtain estimates and hypothesis tests for coefficients corresponding to each pathway. We will also investigate the appropriateness of several structural equations models for the NAS outcomes. First, we will fit a model that considers the joint effects of short- and long-term exposure on EKG pattern via blood pressure and an alternative pathway. In this model, we will specify short-term exposure as having a direct influence on BP. Results will allow us to distinguish between short- and long-term exposure, and potentially suggest that observed EKG changes can be attributed to effects on BP, with other pathways contributing little. We have successfully used this approach to separate out the effects of recent and longer-term exposure in

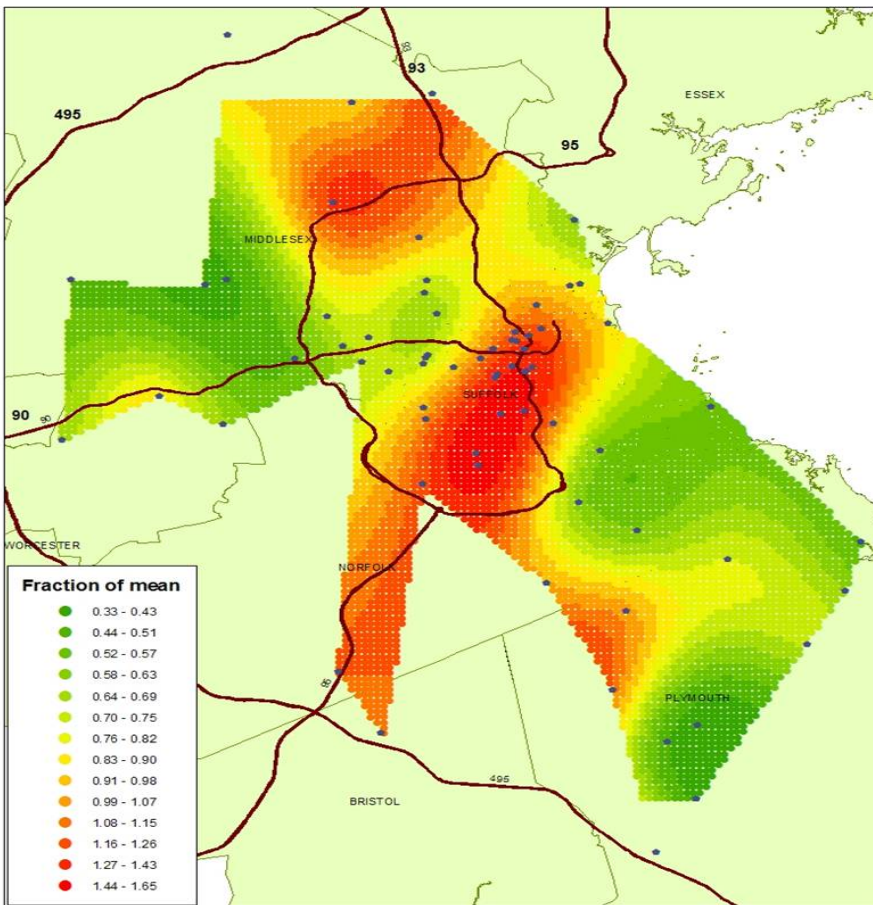
a study of health effects of lead exposure<sup>53</sup>. We will also fit a structural equation model that specifies an association between air pollution exposure and multiple blood markers thought to represent systemic inflammation. In particular, we will fit a model that specifies a single latent variable underlying the multiple outcomes CRP, ICAM-1, VCAM-1, and homocysteine, and estimate the association between PM exposure and this latent variable. Because these outcomes are considered surrogates for systemic inflammation, the model allows us to avoid the multiple testing involved in fitting a model to each outcome separately, while gaining power by pooling information across multiple surrogates of the same underlying state. Finally, we will apply models that consider different potential mechanisms (vascular, autonomic, and inflammatory) simultaneously. This model will allow us to investigate whether pollution is associated with one outcome (say vascular) via another (say, autonomic effects). All of these structural equations models will contain multiple surrogates of exposure to reduce the effects of exposure measurement error. For example, both Black Carbon and particle number are dominated by mobile source emissions, and can be considered surrogates for a latent variable defined as traffic exposure. By assuming that these surrogates reflect some underlying, latent exposure, we assume a factor model which partitions the variance-covariance matrix of the exposures into that explained by the common factor and that explained by measurement error associated with each surrogate. The latent biologic variables are subsequently modeled as a function of the common exposure, thereby adjusting out the measurement error associated with the individual exposure measures. Because such error in multi-pollutant studies typically biases effect estimates downward, and on rare occasions biases them upward, such adjustment may be important in assessing possible health effects associated with particles. We will fit all structural equations models with the commercial software M-plus<sup>54</sup>.

**4.3.2. Spatial-Temporal Modeling:** An important component of this Center proposal is the focus on assessing health effects of different sources of pollution. In the NAS study, health effects will be observed on subjects living throughout Eastern Massachusetts. Because different chemical species have different spatial distributions, with regional pollutants being more homogeneous over space and local sources demonstrating higher spatial variability, investigation of the local pollutant spatial variability can be an important strategy for separating health effects from the different sources. To that end, in addition to baseline analyses that use measurements from central-site and indoor/personal monitors, we will use spatial-temporal modeling of spatially-varying exposures, such as traffic pollution. We will then geocode the location of the study participants' residence, and use exposure estimates as predictors in the health effects analyses.

This approach will yield more accurate exposure measures of spatially heterogeneous pollutants than central site monitoring, and will allow assessment of chronic exposures. This approach can decrease the amount of measurement error associated with the central-site measurements and in turn yield more powerful tests of health effects. To build our model, we will use data from existing individual exposure studies conducted as part of the current EPA PM Center and data from approximately 20 ambient monitoring sites located in the greater Boston area. Collectively, we currently have data from approximately 100 Boston locations during the time period from April 1999 to present. The data currently consist of over 4,300 observations from more than 1,600 unique exposure days. Using these data, we will build and validate models for traffic exposures. Predictions will be based on meteorological conditions and other characteristics of a particular day (e.g. weekday/weekend), as well as measures of the amount of traffic activity (e.g.

GIS-based measures of cumulative traffic density within 100 meters, population density, distance to nearest major roadway, percent urbanization, etc.) at a given location. We will use the readings from the HSPH central site (described in the Particle Technology and Monitoring Core, section 4.1.1) as a predictor. We will use the nonparametric regression methods outlined in Section 3.1.4 to allow these factors to affect exposure levels in a potentially nonlinear way. Finally, we will use a two-dimensional extension of nonparametric regression terms, thin-plate splines, as a function of longitude and latitude to estimate additional spatial variability unaccounted for after including all relevant spatial predictors in the model. Taken together, the model represents a ge additive model as outlined by Kammann and Wand<sup>49</sup>. As an example of this approach, we have done preliminary modeling investigating the suitability of such a model for black carbon levels in Boston. Let  $Y_{ij}$  be the log-transformed black carbon concentration for the  $j^{\text{th}}$  location on day  $i$ . The best model based on the above set of predictors expresses black carbon concentration as a function of year, weekday, black carbon (BC) levels from the HSPH monitoring site, meteorological conditions, cumulative traffic density within one-hundred meters, and a nonparametric function of longitude and latitude. The model is:

$$Y_{ij} = \beta_0 + \beta_1 \text{Year}2000_i + \dots + \beta_4 \text{Year}2003_i + \beta_5 \text{Monday}_i + \dots + \beta_{11} \text{Saturday}_i + \beta_{12} \text{bc.HSPH}_i + f_1(\text{Temp}_i) + f_2(\text{R.Hum}_i) + f_3(\text{Year}day_i) + f_4(\text{Wind}_i) + f_5(\text{C.ADT}_i) + g(\text{lat}, \text{long})_j + \varepsilon_{ij} \quad (7)$$



Results suggest that there exists significant spatial variability in these concentrations in the Boston area.

Figure 1 shows the estimated residual surface as a function of longitude and latitude over the study area. The figure shows that, as expected, the concentrations are highest in the downtown Boston area as well as in the industrial corridor to the northwest. The range of variability is over a factor of three. Note that this spatial surface is after controlling for local traffic counts. Preliminary results also suggest that we have enough data to refine the model to obtain season-specific spatial

**Figure 1: Estimated Spatial Effect in Boston Area BC Levels from Model (7)**

smooths, and that these differ. The adjusted  $R^2$  for this model was 0.81, suggesting that it explains the observed data well. Because primary interest focuses on prediction of levels for locations in which no measurements are taken, we investigated the prediction performance of the model by calculating cross-validation correlations between observed and predicted bc levels (rather than focusing on  $R^2$  values). That is, we repeatedly re-fit the model after removing the data from each individual residential location and calculated the correlation between the observed levels and those predicted from the fit obtained after deleting that location. The resulting cross-validation correlation was 0.57, again suggesting that the model can be useful in refining estimates of traffic exposure across the Boston area. We are currently working on improving this fit. For instance, season-specific fits yield a cross-validation correlation of 0.60. This compares favorably to the model that uses only the HSPH central-site monitor to predict location-specific exposures, which results in a correlation of 0.31. In addition to outdoor black carbon concentrations, our spatial database of exposures also contains measurements of other surrogates of traffic exposure, such as  $PM_{2.5}$  and  $NO_2$ . Because locations often have only two of the three surrogates measured (in various combinations), a multipollutant model for all three surrogates will have greater spatial coverage than a univariate strategy that describes the spatial variability of each pollutant separately. As a result, we will build latent variable models for traffic particles using outdoor black carbon and  $NO_2$  to increase our spatial coverage and hence the predictive performance of the models. This latent variable formulation represents a spatial extension of smooth latent variable models for multivariate curve data<sup>55</sup>. It should be noted that the proposed models for exposure do not yield perfect measures of exposure for each study location, but rather improve the naive estimates provided by the central site monitor. As such, an important consideration for the health effects analyses that use these predictions as covariates will be to statistically adjust for the fact that the exposure metrics are now estimated rather than measured. Several approaches for this adjustment exist. One can use a fully Bayesian approach to jointly model the smoothed exposure surface and the health data. However, this approach induces some “feedback” from the health data to inform the exposure modeling, which is not desirable. An alternative approach is to use an error-in-variables approach in which the pollution estimates and standard errors are used to inform a measurement error model. We will employ this approach, which does not allow for health feedback, but does acknowledge the uncertainty in the exposure estimates from the first stage of the model<sup>56</sup>.

**4.3.3. Quantifying Model Uncertainty in Hierarchical Models:** Existing PM epidemiologic analyses have been criticized because multiple sources of uncertainty are involved in obtaining health effect estimates. One key uncertainty is the shape of the concentration–response relationship. Another is estimating how long one would have to wait after reducing pollution before achieving health improvements. That is, are the associations with twenty–year average exposures, which will change slowly, or are they with recent exposures? Existing approaches for addressing these two questions have limitations. For concentration–response curves, they either assume a parametric form for the relationship, or chose the best fitting model from a set of parametric forms. Alternatively, some have attempted to estimate the relation nonparametrically<sup>57</sup>. These approaches treat all cut-points (knots) where the shape of the relationship can change equally, and require a choice of smoothing parameter. Again, this is usually chosen based on some fitting criteria. Unfortunately, alternatives that fit almost as well might have substantially different shapes. While standard methods report uncertainties in parameters or curves given the model that has been chosen, they do not incorporate uncertainties

about the choice of models. Similarly, distributed lag models allow one to examine the issue of latency between exposure and response, as well as cumulative effects<sup>37</sup>. While models exist to examine these relations, they depend on assumptions, such as how long a lag to examine. Again, goodness of fit criterion is often invoked to answer this question. But this provides no assurance that choices that fit almost as well may yield very different results. We propose to explore Bayesian model averaging as a way of addressing these two forms of model uncertainty in assessing the effect of dose and the timing of dose on survival in the Six Cities Study. This approach does not rely upon effect estimates from a single "final" model, which ignores uncertainty associated with model choice and thus can underestimate the variability associated with these effect estimates, but rather, it takes a weighted average of estimates from a range of plausible models. Because the number of person–years in the Six Cities Study is large and a fully Bayesian approach would be computationally prohibitive, we will take a two-stage approach. In the first stage, we will fit a Cox proportional hazards model containing all confounders of interest, plus a dummy variable for each year of follow-up in each city. In the second stage, each candidate model will regress the dummy variables in a linear regression against the annual average PM<sub>2.5</sub> exposure in that city, in that year. We will fit several candidate models that vary according to the shape of the dose-response relationships and the timing in lags at this second stage, and then implement Bayesian model averaging software for linear regression, which is available in Splus<sup>58</sup>. This approach reports the posterior probability of the each candidate model  $M_k$ ,  $k=1,\dots,K$ , which is given by:

$$P(M_k | Y) = \frac{m(Y | M_k)p(M_k)}{\sum [m(Y | M_k)p(M_k)]} \quad (8)$$

where,  $m(Y | M_k) = \int p(Y | \mathcal{G}, M_k)p(\mathcal{G} | M_k)d\mathcal{G}$  is the marginal distribution of the data given the model, and  $p(M_k)$  is the prior probability mass assigned to model  $k$ . The estimated lagged concentration–response curve is simply the weighted average of the estimates from each model, using these posterior probabilities as weights.

**4.3.4. Methods for Linking PM Effects of Pollution Sources:** Methods such as source apportionment and multivariate receptor modeling, specific rotation factor analysis<sup>59</sup>, positive matrix factorization (PMF)<sup>60</sup> and the methods implemented in the EPA program UNMIX<sup>61</sup> are well studied for purposes of investigating sources of particles. However, only a few studies have used these source apportionment techniques to link the different source classes to health effects<sup>62</sup>. In the past we have used a two-stage approach to estimate source-specific health effects. First we performed a source apportionment and then used the resulting calculated source contributions in the health effects model<sup>4,5</sup>. This work, which is supported by the current EPA Center, is straightforward to implement; however, it does not account for the uncertainty in the source composition profiles. Previous statistical research has shown that, in simpler models, measurement error associated with estimated latent variables can lead to bias in the subsequent regression coefficient estimates<sup>63,64,65</sup>. We propose to quantify this bias in the source-apportionment context, and develop methods that account for this uncertainty. Our first step will be to assess the statistical properties of the two-stage estimates of health effects based on estimated source contributions. We will do this by conducting simulations to obtain Monte Carlo estimates of the distributions of the health effects estimates for a known mixture of source-specific pollution and known health effect. In order to make our findings most relevant to Harvard research interests, we will use a large existing PM Center dataset of Boston daily

elemental concentrations to obtain reasonable settings for all parameters in the simulation. For each simulated dataset, we will apply the two-stage approach as well as the tracer approach (See Section 3.1.1) and assess the performance of each. We will also derive analytical expressions of the bias in the health effects estimates under the two-stage approach, along the lines of the formulas presented by Roberts, Ryan, and Wright<sup>63</sup> for a simpler latent variable setting. Next, we will develop new methodology to estimate the health effects of source-specific components of air pollution based on the full likelihood for exposures and outcome<sup>63,65</sup>. This essentially corresponds to a structural equations approach to the problem. Potential estimation strategies for the health effects in this general framework include maximum likelihood and Bayesian approaches.

## **5. EXPECTED RESULTS**

Many of the tools developed by our Biostatistics group have been essential for our efforts to address key scientific issues pertaining to PM health effects including measurement error, harvesting, identification of susceptible populations, linking between effects and specific sources, and assessing the effects of gaseous co-pollutants. Much of this work has been included in the recent PM Criteria Document and has been of importance in our efforts to assess the validity of epidemiological studies. This Core will continue to apply state-of-the-art analytical tools to our epidemiological, toxicological and exposure data and continue to develop new statistical approaches to investigate PM effects. As in the past, the proposed Center will share these new tools with other researchers in the PM field.

## **6. GENERAL PROJECT INFORMATION**

The Core will be led by Dr. Brent Coull, who will be assisted by Core Co-Leader Dr. Joel Schwartz and Co-Investigator Dr. Antonella Zanobetti. These investigators have worked closely together for over five years and have co-authored many papers, as discussed above. In addition, they have worked closely with the epidemiologists, exposure assessors, and toxicologists of the current Center and these collaborations have produced many interdisciplinary peer-reviewed publications.

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.

## 7. REFERENCES

1. Batalha, JRF, Saldiva, PHN., Clarke, RW, Coull, BA, Stearns, RC, Lawrence, J., Krishna Murthy, GG, Koutrakis, P, and Godleski, JJ. Concentrated ambient air particles induce vasoconstriction of small pulmonary arteries in rats. *Environmental Health Perspectives* 2002; 110:1191-1197.
2. Saldiva, P, Clarke, RW, Stearns, R, Lawrence, J, Koutrakis, P, Hsu, H, Tsuda, A, Coull, BA, Godleski, JJ. The acute inflammatory reaction induced by concentrated ambient particles is influenced by particle composition. *Journal of Respiratory and Critical Care Medicine* 2002; 165:1610-1617.
3. Wellenius, GA, Coull, BA, Godleski, JJ, Koutrakis, P, Okabe, K., Savage, ST, Lawrence, JE, Krishna Murthy, K, Verrier, RL. Inhalation of concentrated ambient air particles exacerbates myocardial ischemia in conscious dogs. *Environmental Health Perspectives* 2003; 111:402-408.
4. Clarke, RW, Coull, BA, Reinisch, U, Catalano, PJ, Killingsworth, CR, Koutrakis, P, Kavouras, I., Lawrence, J, Lovett, E, Wolfson, JM, Verrier, RL, and Godleski, JJ. Inhaled concentrated ambient particles induce pulmonary inflammation and hematological changes in canines. *Environmental Health Perspectives* 2000; 12:1179-1187.
5. Laden, F, Neas, LM, Dockery, DW, Schwartz, J. Association of fine particulate matter from different sources with daily mortality in six U.S. cities. *Environmental Health Perspectives*, 2000; 108:941-7.
6. Long, CM, Suh HH, Catalano PH, Koutrakis, P. Using time- and size-resolved particulate data to quantify indoor penetration and deposition behavior. *Environmental Science & Technology* 2001; 35:2089-2099.
7. Sarnat JA, Schwartz J, Catalano PJ, Suh HH. Gaseous pollutants in particulate matter epidemiology: Confounders or surrogates? *Environmental Health Perspectives* 2001; 109:1053-1061.
8. Sarnat, JA, Long, CM, Koutrakis, P, Coull, BA, Schwartz, J, Suh, HH. Using Sulfur as a Tracer of Outdoor Fine Particulate Matter. *Environmental Science & Technology* 2002; 36:5305-5314.
9. Wellenius, GA, Batalha, JRF, Lawrence, J, Coull, BA, Katz, T, Diaz, EA, Verrier, RL, Godleski, JJ. Effects of carbon monoxide and ambient particles on the incidence of ventricular arrhythmias in a rat model of myocardial infarction. *Toxicological Sciences* 2004; 80:367-376.
10. Gold, DR, Litonjua, A., Schwartz, J, Lovett, E, Larson, A, Nearing, B, Allen, G, Verrier, M, Cherry R, Verrier, R. Ambient pollution and heart rate variability. *Circulation* 2000; 1010:1267-73.
11. Gold, DR, Litonjua, A, Zanobetti, A, Coull, BA, Schwartz, J, MacCallum, G, Verrier, R, Nearing, B, Jacobson, M, Suh, H, and Stone, P. Traffic pollution and ST-segment depression: the relative importance of black carbon and carbon monoxide. 2004. Under revision for publication in *Circulation*.
12. Braga, ALF, Zanobetti, A, Schwartz, J. The Lag Structure Between Particulate Air Pollution and Respiratory and Cardiovascular Deaths in Ten US Cities. *Journal of Occupational and Environmental Medicine* 2000; 43:927-933.

13. Braga, ALF, Zanobetti, A, Schwartz J. The Time Course of Weather Related Deaths. *Epidemiology* 2001; 12:662-667.
14. Schwartz, J and Coull, BA. Control for confounding in the presence of measurement error in hierarchical models. *Biostatistics* 2003; 4:539-553.
15. Schwartz, J, Laden, F, Zanobetti, A. The concentration-response relation between pm2.5 and daily deaths. *Environmental Health Perspectives* 2002; 110:1025-1029.
16. Fitzmaurice, GM, Laird, LM, Ware, JH. *Applied Longitudinal Analysis*. 2004, New York: Wiley.
17. Verbeke, G and Molenberghs, G. *Linear Mixed Models for Longitudinal Data*. 2000. New York: Springer-Verlag.
18. Schwartz, J. Air pollution and daily mortality in Birmingham, Alabama. *American Journal of Epidemiology* 1993; 137:1136-47.
19. Coull, BA, Catalano, PJ, Godleski, JJ. Semiparametric analyses of cross-over data with repeated measures. *Journal of Agricultural, Biological, and Environmental Statistics* 2000; 5:417-429.
20. Godleski, JJ, Verrier, RL, Koutrakis, P, Catalano, PJ, Coull, BA, Reinisch, U, Lovett, EG, Lawrence, J, Krishna Murthy, GG, Wolfson, JM, Clarke, RW, and Nearing, BD. (2000). "Mechanisms of Morbidity and Mortality from Exposure to Ambient Air Particulate." *Health Effects Institute Research Report* 2000; 91:1-103.
21. Hastie, T. and Tibshirani, R. *Generalized additive models*. 1990, London: Chapman and Hall.
22. Dominici F, McDermott A, Zeger SL, Samet JM. On the use of generalized additive models in time-series studies of air pollution and health. *American Journal of Epidemiology* 2002; 156:193-203.
23. Eilers, PHC and Marx, BD. Flexible smoothing with B-splines and penalties (with discussion). *Statistical Science* 1996; 89:89-121.
24. Ma RJ, Krewski D, Burnett RT. Random effects Cox models: A Poisson modelling approach. *Biometrika* 2003; 90:157-169.
25. Schwartz, J. and Zanobetti, A. Using Meta-Smoothing to Estimate Dose-Response Trends across Multiple Studies, with Application to Air Pollution and Daily Death. *Epidemiology* 2000; 11:666-672.
26. Berkey, CS, Hoaglin DC, Mosteller, F, Colditz, GA. A random-effects regression model for meta- analysis. *Statistics in Medicine* 1995; 14:395-411.
27. Sarnat, JA, Brown, KW, Schwartz, J, Coull, BA, Koutrakis, P. Relationships among Personal Exposures and Ambient Concentrations of Particulate and Gaseous Pollutants and their Implications for Particle Health Effects Studies. *Epidemiology* 2004, in press.
28. Zeka, A and Schwartz, J. Estimating the independent effects of multiple air pollutants in the presence of measurement error: an application of a measurement error resistant technique. 2004. Submitted for publication.
29. Koutrakis, P, Suh, HH, Sarnat, JA, Brown, KW, Coull, BA, Schwartz, J. Characterization of Particulate and Gas Exposures of Sensitive Subpopulations Living in Baltimore and Boston. *Health Effects Institute Research Report* 2004, in press.
30. Coull, BA, Schwartz, J, Wand, MP. Respiratory Health and Air Pollution: Additive Mixed Model Analyses. *Biostatistics* 2001; 2:337-349.

31. Brumback, BA, Ruppert, D, Wand, MP. Comment to "Variable selection and function estimation in additive nonparametric regression using data-based prior". *Journal of the American Statistical Association* 1999; 94:794-797.
32. Wand, MP, Coull, BA, French, JL, Ganguli, B., Kammann, EE, Staudenmayer, J., Zanobetti, A. (2003). Semipar: A Module for Semiparametric Regression in Splus. Release 1.0. Available at <http://www.maths.unsw.edu.au/~wand/semipar.html>.
33. Zhao, Y, Staudenmayer, J, Coull, BA, Wand, MP. General design Bayesian generalized linear mixed models. 2004. Submitted for publication. Available at <http://www.maths.unsw.edu.au/~wand/papers.html>.
34. Zanobetti, A, Wand, MP, Schwartz, J, Ryan, LM. Generalized additive distributed lag models: quantifying mortality displacement. *Biostatistics* 2000; 1: 279-292.
35. Goodman, PG, Dockery, DW, Clancy, L. Cause-specific mortality and the extended effects of particulate pollution and temperature exposure. *Environmental Health Perspectives* 2004; 112:179-185.
36. Schwartz, J. The distributed lag between air pollution and daily deaths. *Epidemiology* 2000; 11:320-326.
37. Zanobetti, A, Schwartz, J, Gryparis, A, Touloumi, G, Atkinson, R, Le Tertre, A, Bobros, J, Celko, M, Goren, A, Forsberg, B, Michelozzi, P, Rabczenko, D, Aranguiz Ruiz, E, Katsouyanni, K. The temporal pattern of mortality responses to air pollution. *Epidemiology* 2002; 13:87-93.
38. Zanobetti A, Schwartz J, Samoli E, Gryparis A, Touloumi G, Peacock J, Anderson RH, Le Tertre A, Bobros J, Celko M, Goren A, Forsberg B, Michelozzi P, Rabczenko D, Hoyos SP, Wichmann HE, Katsouyanni K. The temporal pattern of respiratory and heart disease mortality in response to air pollution. *Environmental Health Perspectives* 2003; 111:1188-93
39. Maclure, M. The case-crossover design: a method for studying transient effects on the risk of acute events. *American Journal of Epidemiology* 1991; 133:144-53.
40. Bateson, T and Schwartz, J. Selection Bias and Confounding in Case-Crossover Analyses of Environmental Time Series Data. *Epidemiology* 2001; 12:654-661.
41. Bateson, T.F. and Schwartz, J. Who is sensitive to the effects of particles on mortality? A case-crossover analysis of individual characteristics as effect modifiers. *Epidemiology* 2004; 15:143-149.
42. Lumley, T. and Levy, D. Bias in the case-crossover design: implications for studies of air pollution. *Environmetrics* 2000; 11:689-704.
43. Schwartz, J. Is the association of airbourne particles with daily deaths confounded by gaseous air pollutants: an approach to control by matching. *Environmental Health Perspectives* 2004; 112:557-61.
44. Coull, BA, Hobert, JP, Ryan, LM, Holmes, LB. Crossed random effect models for multiple outcomes in a study of teratogenesis. *Journal of the American Statistical Association* 2001; 96:1194-1204.
45. Shen W and Louis, TA Triple-goal estimates in two-stage hierarchical models. *Journal of the Royal Statistical Society, Series B* 1998; 60:455-471.
46. Houseman, EA, Coull, BA, Ryan, LM. A functional-based distribution diagnostic for a linear model with correlated outcomes. 2004. Under revision for publication in *Biometrika*.

47. Houseman, EA, Ryan, LM, Coull, BA. Cholesky residuals for assessing normal errors in a linear model with correlated outcomes. *Journal of the American Statistical Association* 2004; 99:383-394.
48. Agresti, A. *Categorical Data Analysis*. (2<sup>nd</sup> Ed.). 2004. New York: Wiley.
49. Kammann, EE and Wand, MP. Geoadditive models. *Journal of the Royal Statistical Society, Series C*, 2003; 52:1-18.
50. Ruppert, D, Wand, MP, Carroll, RJ. *Semiparametric Regression*. 2003. New York: Chapman & Hall.
51. Schwartz, J., Societal benefits of reducing lead exposure. *Environmental Research* 1994; 66:105-124.
52. Budtz-Jorgensen E, Keiding N, Grandjean P, Weihe P, White RF. Statistical methods for the evaluation of health effects of prenatal mercury exposure. *Environmetrics* 2003; 14:105-112.
53. Chuang HY, Schwartz J, Gonzales-Cossio T, Lugo MC, Palazuelos E, Aro A, Hu H, Hernandez-Avila M. Interrelations of lead levels in bone, venous blood, and umbilical cord blood with exogenous lead exposure through maternal plasma lead in peripartum women. *Environmental Health Perspectives* 2001; 109:527-532
54. Muthen, LK and Muthen, BO. *Mplus User's Guide*. (3<sup>rd</sup> Ed.). 1998. Los Angeles, CA: Muthen & Muthen.
55. Coull, BA. and Staudenmayer, J. Self-modeling regression for multivariate curve data. *Statistica Sinica* 2004; 14:695-712.
56. Shaddick, G and Wakefield, J. *Modelling* daily multivariate pollutant data at multiple sites. *Journal of the Royal Statistical Society, Series C* 2002; 51:351-372.
57. Schwartz, J. The use of epidemiology in environmental risk assessment. *Journal of Human and Ecological Risk Assessment* 2002; 8:1253-1265.
58. Raftery, AE, Madigan, D, Hoeting, JA. Bayesian modeling averaging for linear regression models. *Journal of the American Statistical Association* 1997; 92:179-191.
59. Koutrakis, P and Spengler, JD. Source apportionment of ambient particles in Steubenville, Ohio using specific rotation factor analysis. *Atmospheric Environment* 1987; 21:1511-1519.
60. Paatero, P and Tapper, U. Positive matrix factorization – A nonnegative factor model with optimal utilization of error-estimates of data values. *Environmetrics* 1994; 5:111-126.
61. Willis, DA. Workshop on UNMIX and PMF as Applied to PM<sub>2.5</sub>. 2000. Research Triangle Park: EPA.
62. Lumley, T and Liu, H. How can source apportionment and receptor modeling data be used in epidemiology? Presentation at the American Association for Aerosol Research Conference "Receptor Modeling and Source Apportionment". Pittsburgh, April 1, 2003.
63. Roberts, K, Ryan, LM, Wright, RJ. On the use of Rasch models for handling high-dimensional covariates in epidemiological studies. 2003. Under revision for publication in *Biometrics*.
64. Tsiatis, AA, De Gruttola, V, Wulfsohn, MS. Modeling the relationship of survival to longitudinal data measured with error; applications to survival and CD4 counts in patients with AIDS. *Journal of the American Statistical Association* 1995; 90:27-37.

65. Zhang, D and Lin, X. Generalized linear models with longitudinal covariates. Technical Report. Department of Statistics, North Carolina State University. 1999. Available at <http://www4.stat.ncsu.edu/~dzhang2/method.html>.