JSM 2012: Section on Risk Analysis
Student/Young Researcher Paper Competition

Using Semiparametric Mixed Model and Functional Linear Model to Detect Vulnerable Prenatal Window to Carcinogenic Polycyclic Aromatic Hydrocarbons on Fetal Growth

Lu Wang · Xihong Lin · Hyunok Choi

Abstract Prenatal exposure to carcinogenic polycyclic aromatic hydrocarbons (c-PAHs) through maternal inhalation induces higher risk for a wide range of fetotoxic effects. However, the question of whether there is a gestational window during which the human embryo/fetus is particularly vulnerable to PAHs has not been examined thoroughly. We consider a longitudinal semiparametric mixed effect model to characterize the individual prenatal PAH exposure trajectory. Parametric fixed effects are used to model the covariate effects and a fully nonparametric smooth function is used to model the time effect. The within-subject correlation is modeled using subject-specific random effects. We maximize the penalized likelihood to estimate the regression coefficients and the nonparametric function of time, whose estimator is a cubic smoothing spline. The smoothing parameter, along with the variance components, is simultaneously estimated through restricted maximum likelihood estimation by treating the inverse smoothing parameter as an extra variance component in a modified mixed model. The subject specific trajectory of prenatal exposure is estimated using best linear unbiased prediction and is linked to the birth outcomes through a set of functional linear models, where the coefficient of PAH exposure is a fully nonparametric function of gestational age. This allows the effect of PAH exposure on each birth outcome to vary at different gestational age, and the window associated with significant adverse effect is identified as a vulnerable prenatal window to PAHs on fetal growth. We minimize the penalized sum of squared errors using a spline-based expansion of the nonparametric coefficient to draw statistical inferences, and the smoothing parameter is chosen through cross-validation.

Keywords Children's health · Environmental health · Longitudinal study · Risk assessment · Spline basis · Windows of vulnerability

Lu Wang
Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA
E-mail: luwang@umich.edu

Xihong Lin
Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, USA

Hyunok Choi
Department of Environmental Health Sciences, University at Albany, Rensselaer, NY 12144, USA

1 Introduction

Within areas of environmental health research, especially concerning children's health, there has been an increasing interest in identifying the periods when exposure to environmental toxicants causes a higher risk or a stronger health deficit later in life compared to other periods when the exposure occurs (Selevan et al. 2000; Barr et al. 2000; West 2002). For example, prenatal or early postnatal exposure to polycyclic aromatic hydrocarbons (PAHs), which are emitted during incomplete combustion and/or pyrolosis of fossil fuel, coal, wood, cigarette and food items, exerts both developmental toxicity, carcinogenicity and disruption of the endocrine system (Castro et al. 2008; Perera et al. 2005; Yu et al. 2006). Strong associations of prenatal exposure to PAHs with small-for-gestational age, preterm delivery, and neuro-developmental deficits in children have also been observed (Schwartz 2004; Kim 2004; Selevan et al. 2000). However, the question of whether there is a gestational window during which the human embryo/fetus is subject to a higher risk if exposed to PAHs has not been examined thoroughly. It is of great clinical importance to identify such critical windows of vulnerability, and thus avoid unnecessary toxic exposures to reduce the risk of intrauterine growth restriction (IUGR). In public health and risk management, information on such critical windows of vulnerability may also help identify specific interventions for susceptible subgroups.

Early-life exposure to xenotoxins, spanning from embryo to early childhood, is of particular interest not only because it is the period of exquisite vulnerability, but also because of its possible 'programming' role in immune, metabolic, and neurological functions throughout the life course. This unique vulnerability of the fetus to xenotoxins has been attributed to susceptibility to epigenetic disruption, immaturity of immune systems, the rapid development of fetal organs, and the fact that exposure per body weight is much higher than that for adults. Furthermore, maternal milieu varies over the pregnancy period, with changes in absorption, distribution, metabolism, and excretion of xenobiotics. Subtle morphological and/or functional modifications due to prenatal xenotoxin exposure have been associated with an increased risk of many illnesses, including delayed cognitive function, cardiopulmonary diseases, diabetes during adulthood, lymphoma, breast cancer, and Parkinson's disease.

Question of whether the timing of prenatal exposure to airborne PAHs induces variable risks of IUGR has garnered a considerable interest (Sram 2005; Sanyal and Li 2007). Traditional consensus has been that the largest fetal weight gain is during the third trimester and accordingly xenotoxin exposure in the later half-of the pregnancy period has been deemed the most detrimental. However, robust body of experimental data demonstrates that PAH exposure during the earliest gestational weeks profoundly affects the subsequent disease risks. Exposure to PAHs, in particular B[a]P and 7,12-dimethylbenz[a]anthracene (DMBA) during organogenesis leads to significant reduction in birth

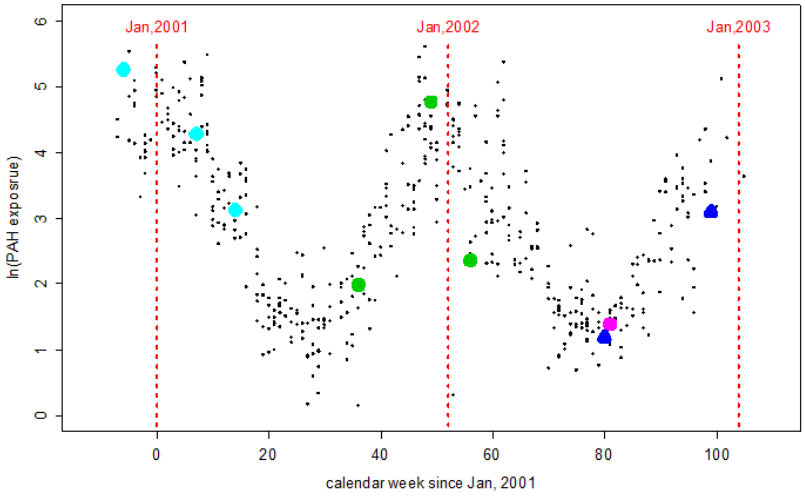


Fig. 1: Personal prenatal PAH exposure measurements across the calendar time among the study cohort from November 2000 to January 2003. The observations of four subjects are illustrated using different subject-specific colors.

weight, crown-lump length, as well as placental proficiency. Fetal cranium and fetal neural tissues appear to be particularly sensitive to B[a]P and DMBA exposure. Therefore, transplacental exposure to B[a]P significantly impairs long-term potentiation, a marker of long-term memory and learning. In humans, it remains unclear whether the gestational age-specific PAH exposure differentially affects functions and/or physiology of the developing systems. While perinatal period is generally regarded as the most susceptible period during human development, functional and/or physiological alteration due to PAH exposure during a narrower and more precise vulnerable gestational window remains very poorly understood. In order to better understand the etiologies underlying the intrauterine growth restriction, we focus here on the question of variability in intrauterine growth restriction risk associated with comparable unit of exposure across different gestation age.

In this paper, we propose a novel modeling procedure to identify the critical vulnerable prenatal window to PAH on fetal growth, based on a prospective cohort study with unique personal PAH exposure measurements conducted in Krakow, Poland. (Jedrychowski et al. 2003, 2004, 2006; Choi et al. 2006, 2008). To investigate the effect of prenatal and early childhood exposure to multiple toxicants on a number of developmental and health outcomes, a cohort of pregnant and healthy women was enrolled in Krakow between 2000 and 2003. During pregnancy, a questionnaire on health history, lifestyle, and home environment was administered repeatedly. The participants were also invited to undergo a 48-hr personal air monitoring to estimate their personal prenatal exposure to PAHs during each trimester. Figure 1 shows the exposure measurements from all subjects during the study period and suggests that the prenatal exposure level varies over time in a complicated periodic manner. Thus, modeling its time trend using a simple parametric function would be difficult. Also, the personal measurements are sparse

due to technical difficulty, pregnancy burden, and expense considerations. Therefore, nonparametric modeling is not only more flexible in the functional form, but also enables us to borrow the information from others to help estimate subject-specific exposure trajectories. We estimate individual prenatal exposure curve for the entire duration of pregnancy through a longitudinal semiparametric mixed effect model (Zhang et al. 1998; Ke and Wang 2001; Elmi et al. 2011). Parametric fixed effects are used to model the covariate effects and a fully nonparametric smooth function is used to model the time effect. The within-subject correlation is modeled using subject-specific random effects. We use maximum penalized likelihood to estimate the regression coefficients and the nonparametric function of time, whose estimator is shown to be a cubic smoothing spline. We use restricted maximum likelihood to simultaneously estimate the smoothing parameter and the variance components. Using cubic spline basis, all model parameters can be obtained by fitting a modified linear mixed model with the inverse smoothing parameter as an extra variance component. With the predicted subject-specific prenatal PAH exposure trajectory, we propose a set of functional linear models (Ramsay and Silverman 1997; Ramsay et al. 2009; Cardot 2003) to relate the birth outcomes and the individual specific curves of prenatal PAH exposures. The coefficient associated with PAH exposure in functional linear models is a fully nonparametric function of gestational age, which allows the effect of PAH exposure on each birth outcome to vary at different gestational age, as well as the corresponding significance. The gestational window associated with significantly detrimental effects is identified as a vulnerable window to PAHs on fetal growth. We employ a spline-based expansion of the nonparametric coefficient curve and minimize the penalized sum of squared errors to draw statistical inferences. The smoothing parameter is chosen through cross-validation.

2 The Statistical Models

Eight carcinogenic PAHs were monitored and measured during the unique 48-hr personal air monitoring in the study. We use the sum of eight c-PAHs to summarize the PAH exposure level at each observation. A log transformation is applied to the PAH level in order to get a more normally distributed measurement. To account for the correlation between the multiple monitoring of the same subject at different trimesters, and to account for the nonlinear effect of calendar time on the PAH exposure level as shown in Figure 1, we fit a semiparametric mixed effect model of log PAH with a random intercept and a random slope as

$$log(Y_{ij}) = \boldsymbol{X}_{ij}^T\boldsymbol{\beta} + f(t_{ij}) + b_{i1} + b_{i2}\cdot t_{ij} + e_{ij}, \tag{1}$$

where Y_{ij} denotes the jth PAH exposure measurement for subject i at time t_{ij}, $i = 1, \cdots, n$, $j = 1, \cdots, n_i$, n_i denotes the number of PAH measurements for subject i, which may vary from one subject to another, and $\boldsymbol{\beta}$ is a $d_1 \times 1$ vector of regression coefficients associated with the covariates $\boldsymbol{X}_{ij}$. We model

the effect of calender time t_{ij} on the log PAH exposure level through a fully nonparametric function $f(t_{ij})$ to allow for any possible non-linear association. $f(t)$ is assumed to be a twice-differentiable smooth function. The effect of other potential predictors, such as spatial factors (whether living in City center) and behavioral factors (whether smoke), are modeled linearly through $\boldsymbol{X}_{ij}^T\boldsymbol{\beta}$ in the model. We use b_{i1} and b_{i2}, the random intercept and random slope for the ith subject respectively, to handle the within-subject correlation. We assume $(b_{i1}, b_{i2})^T \sim N(\mathbf{0}, \boldsymbol{\Sigma})$. $e_{ij} \sim N(0, \sigma_e^2)$ are independent measurement errors. Using such a semiparametric model, each subject borrows information from the other subject in the study cohort to fit the whole personal PAH exposure profile. By including both random intercept and random slope, the predicted individual PAH exposure trajectory not only shift from the population mean curve by a subject specific amount b_{i1}, but also has a subject specific different slope with a departure as b_{i2}.

From every subject's estimated individual profile of the PAH exposure over calendar time, we cut out the period from the time she got pregnant to the time she delivered the baby. These individual prenatal PAH exposure curves over their own entire gestational period are put in a functional linear model to assess the association between each birth outcome and individual prenatal PAH exposure. The functional linear model is

$$O_i = \int_0^{T_d} \widehat{Y}_i(t)\alpha(t)dt + \boldsymbol{Z}_i^T\boldsymbol{\gamma} + \epsilon_i, \tag{2}$$

where for subject i, O_i denotes the birth outcome of interest, either birth weight, birth length, or birth circumference, $\widehat{Y}_i(t)$ is the PAH exposure profile predicted from model (1), and $\alpha(t)$ is the coefficient associated with PAH exposure at t. Note t in model (2) denotes the gestational age instead of calendar time and the integral is from gestational age 0 to the gestational age of delivery T_d. $\alpha(t)$ is a fully nonparametric function of t, which allows the effect of PAH exposure on each birth outcome to vary at different gestational age. We controlled for other potential risk factors such as the gestational age, newborn gender, parity, pre-pregnancy weight as well as maternal height through $\boldsymbol{Z}_i^T\boldsymbol{\gamma}$, where $\boldsymbol{Z}_i$ denote these potential risk factors for poor birth outcome and $\boldsymbol{\gamma}$ is a $d_2 \times 1$ vector of regression coefficients associated with the covariates $\boldsymbol{Z}_i$. We assume $\epsilon_i \sim N(0, \sigma_\epsilon^2)$ are independent random errors.

3 Estimation Procedure

3.1 Penalized Likelihood Estimation in Semiparametric Mixed Effect Model

To facilitate the presentation, we introduce the following matrix notation. Let $\boldsymbol{Y}_i = (Y_{i1}, \cdots, Y_{in_i})^T$, for $i = 1, \cdots n$, and similarly define $\boldsymbol{X}_i$ and $\boldsymbol{e}_i$. Also, define $\boldsymbol{Z}_i = (\mathbf{1}_{n_i\times 1}, \boldsymbol{t}_i)$, where $\mathbf{1}_{n_i\times 1} = (1\cdots 1)_{n_i\times 1}^T$ and $\boldsymbol{t}_i = (t_{i1}, \cdots, t_{in_i})^T$ are both $n_i \times 1$ vectors. Recall t_{ij} denotes the calendar time when PAH is measured the jth time for subject i. Let $\widetilde{\boldsymbol{t}} = \left(\widetilde{t}_1, \cdots, \widetilde{t}_r\right)^T$ denote a $r \times 1$ vector of ordered distinct

values of the observed measure time points $\{t_{ij}: \ i = 1, \cdots, n, \text{ and } j = 1, \cdots, n_i\}$. We define $\boldsymbol{M}_i$ as an incidence matrix for the ith subject to connect the observed measure time for subject i and $\widetilde{\boldsymbol{t}}$. The (j,l)th element of $\boldsymbol{M}_i$ is 1 if $t_{ij} = \widetilde{t}_l$ and 0 otherwise, where $j = 1, \cdots, n_i, \ l = 1, \cdots r$. Then model (1) can be re-written using matrix notation as

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{M}_i\boldsymbol{f} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{e}_i$$

where $\boldsymbol{f} = \left\{f\left(\widetilde{t}_1\right), \cdots, f\left(\widetilde{t}_r\right)\right\}^T$, and $\boldsymbol{b}_i = (b_{i1}, b_{i2})^T$. Considering all the subjects in the study and further denoting $\boldsymbol{Y} = \left(\boldsymbol{Y}_1^T, \cdots, \boldsymbol{Y}_n^T\right)^T$, $\boldsymbol{X} = \left(\boldsymbol{X}_1^T, \cdots, \boldsymbol{X}_n^T\right)^T$, $\boldsymbol{M} = \left(\boldsymbol{M}_1^T, \cdots, \boldsymbol{M}_n^T\right)^T$, $\boldsymbol{Z} = diag\left(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n\right)$, $\boldsymbol{b} = \left(\boldsymbol{b}_1^T, \cdots, \boldsymbol{b}_n^T\right)^T$ and $\boldsymbol{e} = \left(\boldsymbol{e}_1^T, \cdots, \boldsymbol{e}_n^T\right)^T$, we have

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{M}\boldsymbol{f} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{e}, \tag{3}$$

where $\boldsymbol{\beta}$ and $\boldsymbol{f}$ are defined as before, $\boldsymbol{b} \sim N\left\{\boldsymbol{0}, diag\left(\boldsymbol{\Sigma}, \cdots, \boldsymbol{\Sigma}\right)\right\}$, $\boldsymbol{e} \sim N\left\{\boldsymbol{0}, \sigma_e^2\boldsymbol{I}_N\right\}$, and $\boldsymbol{I}_N$ is the identity matrix of dimension $N = \sum_{i=1}^n n_i$. Since $f(t)$ is assumed to be twice differentiable in model (1), we consider a natural cubic spline estimator of $f(t)$ and maximize the penalized likelihood for $\boldsymbol{\beta}$ and $f(t)$ (Zhang et al. 1998; O'Sullivan et al. 1986; Zhang et al. 2007). Specifically in our situation, we maximize

$$l_p\left(\boldsymbol{\beta}, \boldsymbol{f}; \boldsymbol{Y}\right) = l\left(\boldsymbol{\beta}, \boldsymbol{f}; \boldsymbol{Y}\right) - \frac{\lambda}{2}\int_{T_1}^{T_2}\left\{f''(t)\right\}^2 dt.$$

The first term $l\left(\boldsymbol{\beta}, \boldsymbol{f}; \boldsymbol{Y}\right) = \frac{1}{2}\left\{log|\boldsymbol{V}| + \left(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{M}\boldsymbol{f}\right)^T\boldsymbol{V}^{-1}\left(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{M}\boldsymbol{f}\right)\right\}$ is the log-likelihood under model (3), where $\boldsymbol{V} = diag\left(\boldsymbol{V}_1, \cdots, \boldsymbol{V}_n\right)$, and $\boldsymbol{V}_i = \boldsymbol{Z}_i\boldsymbol{\Sigma}\boldsymbol{Z}_i^T + \sigma_e^2\boldsymbol{I}_{n_i}$. The second term is a penalization for the roughness of nonparametric function $f(t)$, where λ is a non-negative smoothing parameter, (T_1, T_2) specifies the range of calendar time t of the study, and $f''(t)$ is the second derivative of $f(t)$. Without loss of generality, we let $T_1 = 0$ by relocating our very first measurement on the calendar time. To facilitate numerical implementation, we define a non-negative definite smoothing matrix $\boldsymbol{K}$ following the equation (2.3) of Green and Silverman (1994), and equivalently maximize

$$l\left(\boldsymbol{\beta}, \boldsymbol{f}; \boldsymbol{Y}\right) - \frac{\lambda}{2}\boldsymbol{f}^T\boldsymbol{K}\boldsymbol{f}. \tag{4}$$

Differentiating (4) with respect to $\boldsymbol{\beta}$ and $\boldsymbol{f}$, we solve

$$\begin{bmatrix} \boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{M} \\ \boldsymbol{M}^T\boldsymbol{V}^{-1}\boldsymbol{X} & \boldsymbol{M}^T\boldsymbol{V}^{-1}\boldsymbol{M} + \lambda\boldsymbol{K} \end{bmatrix}\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{f} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{Y} \\ \boldsymbol{M}^T\boldsymbol{V}^{-1}\boldsymbol{Y} \end{bmatrix}$$

to obtain the estimators $\left(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{f}}\right)$. Subject specific random effects can be calculated using the best linear unbiased prediction (BLUP) estimator

$$\widehat{\boldsymbol{b}}_i = \boldsymbol{\Sigma}\boldsymbol{Z}_i^T\boldsymbol{V}_i^{-1}\left(\boldsymbol{Y}_i - \boldsymbol{X}_i\widehat{\boldsymbol{\beta}} - \boldsymbol{M}_i\widehat{\boldsymbol{f}}\right).$$

Then $\boldsymbol{X}_{i,t}\widehat{\boldsymbol{\beta}} + \hat{f}(t) + \widehat{b}_{i1} + \widehat{b}_{i2}\cdot t$ is the estimated longitudinal trajectory for subject i, where $\boldsymbol{X}_{i,t}$ is the most recent values for covariates at time t. The smoothing parameter λ controls the balance

between the goodness-of-fit of model (1) and the roughness of $f(t)$. Following Zhang et al. (1998), we estimate the smoothing parameter λ by treating its inverse, $1/\lambda$, as an extra variance component in a modified mixed effect model, which is derived from model (3) using the mixed model representation of the smoothing spline estimator of $f(t)$. Specifically, decompose $\boldsymbol{K} = \boldsymbol{L}\boldsymbol{L}^T$, and write $\boldsymbol{f}$ using a mixed effect representation as

$$\boldsymbol{f} = \boldsymbol{A}\boldsymbol{\eta} + \boldsymbol{B}\boldsymbol{d}, \tag{5}$$

where $\boldsymbol{L}$ is a $r \times (r-2)$ full rank matrix, $\boldsymbol{L}^T\boldsymbol{A} = 0$, $\boldsymbol{\eta}$ is a 2×1 fixed effect vector, $\boldsymbol{B} = \boldsymbol{L}\left(\boldsymbol{L}^T\boldsymbol{L}\right)^{-1}$, and $\boldsymbol{d} \sim N\left\{\boldsymbol{0}, \frac{1}{\lambda}\boldsymbol{I}_{(r-2)\times(r-2)}\right\}$ is a $(r-2) \times 1$ random effect vector. Then clearly $\boldsymbol{f}^T\boldsymbol{K}\boldsymbol{f} = \boldsymbol{d}^T\boldsymbol{d}$, and thus the penalized log-likelihood (4) is the same as the log-likelihood of a modified linear mixed model by plugging (5) into (3), with an extra variance component $\frac{1}{\lambda}$ for the additionally introduced random effect $\boldsymbol{d}$. We estimate λ, along with the other variance components simultaneously using the restricted maximum likelihood (REML) approach.

3.2 Estimation in Functional Linear Models Using Spline Basis

Functional linear model (Ramsay and Silverman 1997; Ramsay et al. 2009; Cardot 2003) links a curve predictor to a scalar response variable. For example, in model (2) introduced in Section 2, the response O is a scalar while the predictor $Y(t)$ is a random function of t. The corresponding coefficient $\alpha(t)$ is a nonparametric function. We are faced with the estimation of a functional coefficient or, equivalently, of a linear functional. There have been several approaches proposed in the literature for nonparametric estimation. Geman and Hwang (1982) proposed a sieve maximum likelihood estimation procedure to ease the computational difficulty in fully nonparametric estimation problems. They approximate the unknown nonparametric function by a linear span of some known basis functions to form a sieve log-likelihood. Then maximizing the log-likelihood with respect to the unknown function converts to maximizing the sieve log-likelihood with respect to the finite unknown coefficients in the linear span. Some further theoretical results have been obtained by Shen and Wong (1994). Sieve estimation reduces the dimensionality of the optimization problem, but the number of basis functions also grows as sample size increases. Instead, we consider a regularization approach (Wabba 1990; Ramsay and Silverman 1997), using a similar expansion of $\alpha(t)$ but with fixed number of basis functions. We represent $\alpha(t)$ using spline basis, which has been well recognized in the statistical literature as a useful tool in nonparametric estimation (Stone 1985, 1986). Let us assume $\alpha(t)$ is a smooth nonparametric function which has continuous second order derivative. We minimize the following penalized sum of squared errors

$$\sum_{i=1}^{n}\left\{O_i - \int_0^{T_d}\widehat{Y}_i(t)\alpha(t)dt - \boldsymbol{Z}_i^T\boldsymbol{\gamma}\right\}^2 + \rho\int_0^{T_d}\left\{\alpha''(t)\right\}^2 dt, \tag{6}$$

where $\rho > 0$ is a smoothing parameter to control the roughness penalty. Following Cardot et al. 2003, Ramsay and Silverman 1997, and Ramsay et al. 2009, we consider a B-spline basis $\{S_k(t),\ k = 1, \cdots, K_\alpha\}$, where K_α is the number of basis functions. We represent $\alpha(t)$ as

$$\alpha(t) = \sum_{k=1}^{K_\alpha} \phi_k S_k(t) = \boldsymbol{S}_{K_\alpha}^T(t)\,\boldsymbol{\phi}$$

where ϕ_k are unknown coefficients, $\boldsymbol{S}_{K_\alpha}(t) = \{S_1(t), \cdots, S_{K_\alpha}(t)\}^T$, and $\boldsymbol{\phi} = (\phi_1, \cdots, \phi_{K_\alpha})^T$. Using this representation, model (2) can be re-written as

$$O_i = \sum_{k=1}^{K_\alpha} \phi_k \cdot \int_0^{T_d} \widehat{Y}_i(t) S_k(t)\, dt + \boldsymbol{Z}_i^T \boldsymbol{\gamma} + \epsilon_i.$$

We denote $J_{ik} = \int_0^{T_d} \widehat{Y}_i(t) S_k(t)\, dt$, $\boldsymbol{J}_i = (J_{i1}, \cdots, J_{iK_\alpha})^T$, and then have

$$O_i = \boldsymbol{J}_i^T \boldsymbol{\phi} + \boldsymbol{Z}_i^T \boldsymbol{\gamma} + \epsilon_i.$$

Thus, the penalized residual sum of squares (6) can be re-written as

$$\sum_{i=1}^{n} \left\{O_i - \boldsymbol{J}_i^T \boldsymbol{\phi} - \boldsymbol{Z}_i^T \boldsymbol{\gamma}\right\}^2 + \rho \int_0^{T_d} \left\{\sum_{k=1}^{K_\alpha} \phi_k S_k''(t)\right\}^2 dt,$$

where $S_k''(t)$ is the second derivative of spline function $S_k(t)$. Let $\boldsymbol{U}_{K_\alpha}(t) = \left\{S_1''(t), \cdots, S_{K_\alpha}''(t)\right\}^T$ and let $\boldsymbol{R}$ denote the matrix $\int_0^{T_d} \boldsymbol{U}_{K_\alpha}(t)\, \boldsymbol{U}_{K_\alpha}^T(t)\, dt$, then the last term of the above expression can be simplified as $\rho \boldsymbol{\phi}^T \boldsymbol{R} \boldsymbol{\phi}$. We can further simplify notation by defining $\boldsymbol{\zeta} = \left(\boldsymbol{\phi}^T, \boldsymbol{\gamma}^T\right)^T$, $\boldsymbol{W}_i = \left(\boldsymbol{J}_i^T, \boldsymbol{Z}_i^T\right)^T$, and $\boldsymbol{R}_0$ as a $(K_\alpha + d_2) \times (K_\alpha + d_2)$ matrix, which augments $\boldsymbol{R}$ by attaching $\mathbf{0}$s. Then the penalized objective function (6) that we want to minimize becomes

$$\sum_{i=1}^{n} \left\{O_i - \boldsymbol{W}_i^T \boldsymbol{\zeta}\right\}^2 + \rho \boldsymbol{\zeta}^T \boldsymbol{R}_0 \boldsymbol{\zeta}.$$

We can estimate $\boldsymbol{\zeta}$ by solving

$$\left(\boldsymbol{W}^T \boldsymbol{W} + \rho \boldsymbol{R}_0\right) \boldsymbol{\zeta} = \boldsymbol{W}^T \boldsymbol{O},$$

where $\boldsymbol{O} = (O_1, \cdots, O_n)^T$ is a $n \times 1$ vector, and $\boldsymbol{W} = (\boldsymbol{W}_1, \cdots, \boldsymbol{W}_n)^T$ is a $n \times (K_\alpha + d_2)$ matrix. Correspondingly, we can construct confidence intervals for $\widehat{\boldsymbol{\zeta}}$. We estimate the variance of $\widehat{\boldsymbol{\zeta}}$ as

$$\widehat{Var}\left(\widehat{\boldsymbol{\zeta}}\right) = \widehat{\sigma}_\epsilon^2 \left(\boldsymbol{W}^T \boldsymbol{W} + \rho \boldsymbol{R}_0\right)^{-1} \boldsymbol{W}^T \boldsymbol{W} \left(\boldsymbol{W}^T \boldsymbol{W} + \rho \boldsymbol{R}_0\right)^{-1},$$

and $\widehat{\sigma}_\epsilon^2$ can be calculated from the mean squared residuals. The smoothing parameter ρ is chosen by cross-validation. Let $\widehat{\alpha}_\rho^{(-i)}(t)$ and $\widehat{\boldsymbol{\gamma}}_\rho^{(-i)}$ be the estimates for $\alpha(t)$ and $\boldsymbol{\gamma}$ without the ith observation using smoothing parameter ρ, then the optimal ρ is chosen to minimize

$$CV(\rho) = \sum_{i=1}^{n} \left\{O_i - \int_0^{T_d} \widehat{Y}_i(t) \widehat{\alpha}_\rho^{(-i)}(t)\, dt - \boldsymbol{Z}_i^T \widehat{\boldsymbol{\gamma}}_\rho^{(-i)}\right\}^2.$$

Other versions of cross-validation criteria can be employed in practice to reduce the computational intensity (Ramsay et al. 2009, among others).

4 The Krakow Birth Cohort Study and Results

In the prospective Krakow Birth Cohort Study introduced in Section 1, pregnant women were recruited from prenatal care clinics during their first trimester in Krakow, Poland. In the city of Krakow, coal combustion for domestic heating is the major air pollution source. In contrast, automobile traffic emissions and coal-combustion for industrial activities are relatively minor contributors. We targeted Caucasian pregnant women of ethnic Polish background during the 8th to 13th weeks of gestation. To reduce confounding, only young (age, 18 - 35) and healthy women with no known risks for adverse birth outcomes were eligible. Those who met all the eligibility criteria were simultaneously monitored for their personal (n = 344), home indoor and outdoor exposure levels of PAHs and PM2.5 during the second trimester of pregnancy between November 2000 and January 2003. The women also answered a questionnaire on health, lifestyle and exposure history. In a subset of women, repeated personal monitoring was additionally conducted during the first and the third trimester. The personal exposure measurements in this study are very unique in the literature. Each woman carried or kept near her a personal air monitor which operated for a consecutive 48-hour period. The split flow inlet, placed near the woman's breathing zone, drew in the particulate or semi-volatile vapor PAHs and particles 2.5 m (PM2.5) on a pre-cleaned quartz microfiber filter and polyurethane foam backup. The filters were analyzed for pyrene and eight PAHs known to be carcinogenic as well as having other toxicities: benz(a)anthracene, chrysene/isochrysene, benzo(b)fluoranthene, benzo(k)fluoranthene, benzo(a)pyrene, indeno(1,2,3-cd)pyrene, dibenz(a,h)anthracene and benzo(g,h,i)perylene. We refer to these eight PAHs as carcinogenic PAHs (c-PAHs).

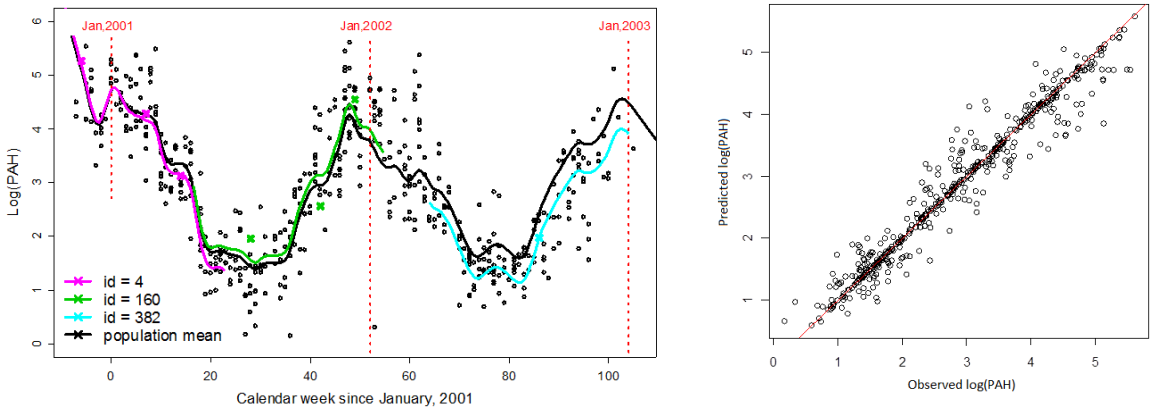
Table 1: Summary of subjects and observations in each monitoring pattern of PAH measurements

	Trimester 1	Trimester 2	Trimester 3	Number of subjects	Number of observations
		X		261	261
			X	5	5
	X	X	X	73	219
	X	X		3	6
		X	X	2	4
	X		X	0	0
Total	**69**	**343**	**83**	**344**	**495**

The personal PAH exposure level is summarized using sum of the eight monitored c-PAHs and a log transformation is taken to make the normality assumption more plausible. The 344 subjects contributed 495 personal PAH measurements from November 2000 to January 2003, while each subject contributed 1 to 3 observations (pattern of observations shown in Table 1). Figure 1 in Section 1 presents the 495 measurements of log PAH exposure over calendar time (in weeks). As illustrated by

several subjects in different colors, the observed measurements of the same subject are very sparse due to technical difficulties related to pregnancy burden and costly expense considerations. Meanwhile, there is a strong evidence in Figure 1 demonstrating a nonlinear periodic effect of calendar time on PAH exposure, which is reasonable given that the study was conducted in about two years. Due to the heating mechanism in Krakow, the PAH exposure level is usually higher in winter and lower in summer. This observable nonlinear periodic effect of the calendar time on the PAH exposure level is captured by our semiparametric model of log PAH. The estimated population mean curve of PAH exposure over calendar time, as well as three individual PAH exposure trajectories over the gestational period that we randomly pick for illustration, are displayed in Figure 2(a).

Fig. 2: Estimation of the population mean curve of PAH exposure among the study cohort and the predictions of individual prenatal PAH exposure during pregnancy based on semiparametric mixed effect model.



(a) Estimation of Exposure Trajectory

(b) Predicted v.s. Observed PAH

As we can see, individual PAH exposure curves estimated from model (1) with a random intercept and a random slope are not just parallel profiles to the estimated population mean. Subject specific trajectories have their own locations and also different slope departures from the estimated population mean. In Figure 2(a), we randomly pick three subjects for illustration. Their observed measurements are reasonably close to the estimated values. This is also confirmed by Figure 2(b), where we plot all observed measurements in the study versus the predicted values, and most points fall along the diagonal line nicely. Model (1) also controls for other potential risk factors. We find that living in City center increases the risk of being exposed to higher PAH level by 0.057 units in log PAH scale (p-value=0.043) while smoking increases the risk of being exposed to higher PAH level by 0.064 units in log PAH scale (p-value=0.01). These results suggest that, in addition to the season effect, host's living condition and personal behavior also critically influences the magnitude of exposure risk to PAH.

Fig. 3: Curve registration: the estimated individual prenatal PAH exposure during pregnancy over gestational age.

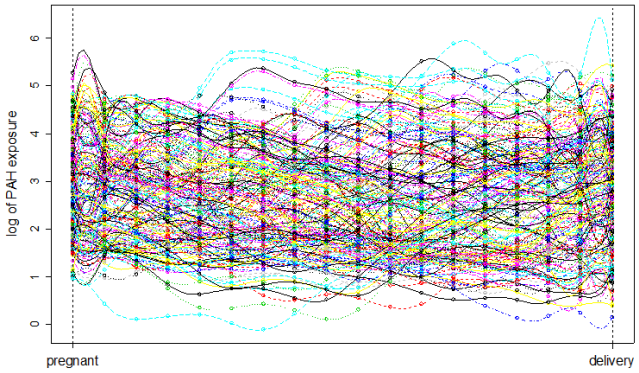
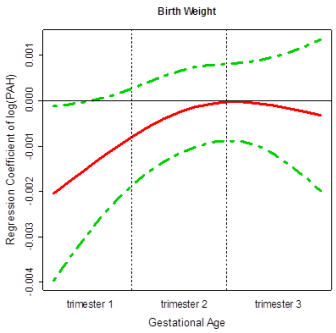
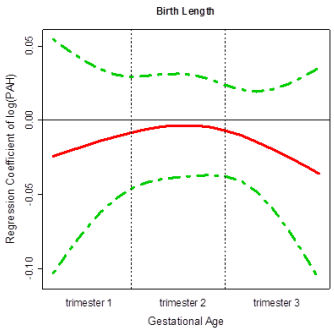


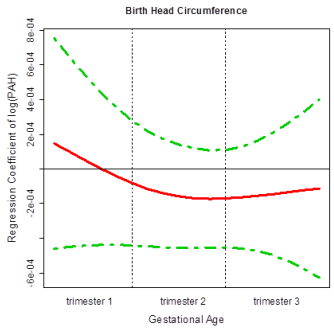
Fig. 4: The estimated regression coefficient $\alpha(t)$ in functional linear model, that is, the effect of prenatal PAH exposure on baby's birth weight, birth length, and birth head circumference across gestational age t. —— is the estimate, and – · – · – is the 95% point-wise confidence interval.



(a) Birth Weight



(b) Birth Length



(c) Birth Head Circumference

When evaluating the risk of intrauterine growth restriction, we are only concerned about the effect of prenatal PAH exposure. Therefore, we cut out the gestational period from the time of fertilization to the time of delivery for each subject on the estimated individual PAH exposure trajectory. This can be viewed as a step of curve registration, where we align the estimated exposure profiles according to gestational age (Figure 3). We preserve the first two trimesters in order to have a biologically meaningful interpretation and only re-scale the last trimester so that the length of gestational period in Figure 3 is the same for everyone. This makes the integral limit in model (2) non-random and facilitates the

estimation of $\alpha(t)$. The aligned and registered curves are used as a predictor in model (2) to assess the effect of prenatal PAH exposure on birth outcomes across different gestational age.

Birth weight, birth length, and birth head circumference are measured for infants at time of delivery. A standard normality test is performed on each birth outcome and log transformation is needed for birth weight. Figure 4 displays the estimated regression coefficient function, $\widehat{\alpha}(t)$, along with 95% point-wise confidence intervals. Obviously, $\widehat{\alpha}(t)$ is not constant for any birth outcome, and thus the risk of intrauterine growth restriction due to prenatal PAH exposure varies across gestational period. Negative values of $\widehat{\alpha}(t)$ means adverse effect, and the magnitude is the loss of baby's birth outcome associated with one unit increase of prenatal PAH exposure at gestational age t. The lower $\widehat{\alpha}(t)$ is, the higher the risk associated with PAH exposure is at this gestational age. For birth weight and birth length, the curves of $\widehat{\alpha}(t)$ are both of bell shape, suggesting that the first and third trimesters are more vulnerable to PAH exposure compared to the second trimester on fetal weight gain and length gain. However, birth head circumference is affected more detrimentally when PAH exposure occurs in the second and third trimesters, which is not surprising given that brain and neural system start to develop rapidly in the second trimester. Only one significant window of vulnerability is identified in the first trimester from fertilization to gestational week 7 for fetal weight gain, where the confidence intervals do not contain zero in Figure 4(a). Additionally, we investigate other potential risk factors including gestational age at born, newborn gender, parity, pre-pregnancy weight, as well as maternal height. The

Table 2: Estimates of the effects of other risk factors in the functional linear models of birth outcomes.

	Birth weight		**Birth length**		**Birth H-C**	
	coefficient	**p-value**	**coefficient**	**p-value**	**Coefficient**	**p-value**
maternal height	0.00013	0.113	0.0024	0.168	3.98e-5	0.158
pre-pregnancy weight	0.0002	<0.001	0.0029	0.009	4.08e-5	0.022
log(gestational age)	0.1305	<0.001	2.1079	<0.001	0.0203	<0.001
parity	0.0023	0.009	0.0256	0.163	0.0011	<0.001
newborn gender	-0.0026	<0.001	-0.0434	0.013	-0.0010	<0.001
c-section delivery					0.00068	0.078

results in Table 2 show that baby boys tend to have a significantly higher birth weight ($p<0.001$), longer birth length ($p=0.013$), and larger birth head circumference ($p<0.001$) than baby girls. Higher pre-pregnancy weight and longer gestational period are significantly associated with an increase of birth weight, birth length, and birth head circumference (all p-values <0.05). Maternal height is positively correlated with all birth outcomes, but none of the associations is statistically significant. Moms who have had baby before tend to bare a baby with higher weight and larger head circumference (p-value are 0.009 and <0.001 respectively). Whether mom has a c-section delivery is also controlled in the model of birth head circumference, but no statistical significance is observed.

5 Concluding Remarks

The identification of a "critical window of vulnerability" to ubiquitous air pollutants such as PAHs is a particularly important, yet challenging question. This is because the dose-response relationship of the xenotoxicant during a given gestational age is inherently related to the host's susceptibility as well as the host's adaptiveness. Furthermore, prenatal exposure to PAHs is chronic throughout the pregnancy period. The concentrations and the relative abundance of PAHs at different gestational age are very likely to vary. However, monitoring the PAH exposure over the whole gestational period is not possible due to technical difficulty and cost considerations. Therefore, statistical methods are in need to provide an efficient and precise estimation of individual prenatal PAH exposure trajectories.

We employ a longitudinal semiparametric mixed effect model to characterize individual profiles of PAH exposure, where the time effect is modeled nonparametrically and random effects are used to account for within-subject correlations. We estimate parametric coefficients and nonparametric function of time by maximizing a penalized likelihood. Subject specific trajectory of PAH exposure is estimated using best linear unbiased prediction. Using curve registration, the prenatal PAH exposure curves are aligned over gestational age and then linked to birth outcomes through functional linear models, where the coefficient of PAH exposure is a fully nonparametric function of gestational age. This allows the effect of PAH exposure on birth outcome to vary at different gestational age. The window associated with significantly high effects is identified as a vulnerable prenatal window to PAHs on fetal growth. To draw statistical inferences in functional linear models, we minimize the penalized sum of squared errors using a spline-based expansion of the nonparametric coefficient function.

Our results show that prenatal PAH exposure is associated with reduction in birth weight, birth length, as well as birth head circumference. The vulnerability of fetus against high prenatal PAH exposure varies across different gestational age. Reducing PAH exposure in both the first trimester and the third trimester are critical for baby's weight and length gain, while birth head circumference is affected more detrimentally when PAH exposure occurs in the second and third trimesters. We detect one critical window of vulnerability in the first trimester from fertilization to gestational week 7, where PAH exposure in this window yields significant impairment on fetal weight gain.

Considering that both proportional and disproportionate intrauterine growth restriction are associated with mortality and morbidity risks of the newborns and compromised cognitive development in children, our data suggest that protection of pregnant women particularly during the first trimester against PAH exposure should be a priority to reduce the risk of intrauterine growth restriction. Ambient PAH concentrations in Krakow are typical of regions dependent on coal-burning for heat and power generation (Junninen et al 2009). The present data support the need for a multinational coal-combustion

abatement strategy for the protection of pregnant women and the embryo/fetus, particularly during the earliest stage of pregnancy.

References

1. Barr M, DeSesso JM, Lau CS, Osmond C, Ozanne SE, Sadler TW, et al. (2000) Workshop to identify critical windows of exposure for childrens health: cardiovascular and endocrine work group summary. Environ Health Perspect, 108: 569–571.
2. Cardot, H., Ferraty, F. and Sarda, P. (1999). Functional linear model. Statist. Probab. Lett. 45: 11–22.
3. Cardot, H., Ferraty, F. and Sarda, P. (2003). Spline estimators for the functional linear model. Statist. Sinica 13: 571–591.
4. Cardot, H. and Sarda, P. (2005). Estimation in generalized linear models for functional data via penalized likelihood, Journal of Multivariate Analysis, 92: 24–41.
5. Castro DJ, Lohr CV, Fischer KA, Pereira CB, and Williams DE (2008) Lymphoma and lung cancer in offspring born to pregnant mice dosed with dibenzo[a,l]pyrene: the importance of in utero vs. lactational exposure. Toxicol Appl Pharmacol, 233: 454–458.
6. Choi H, Jedrychowski W, Spengler J, Camann DE, Whyatt RM, Rauh V, et al. 2006. International studies of prenatal exposure to polycyclic aromatic hydrocarbons and fetal growth. Environ Health Perspect, 114: 1744-1750.
7. Choi, H., Perera, F., Pac, A., Wang, L., Flak, E., Mroz, E., Jacek, R., Chai-Onn, T., Jedrychowski, W., Masters, M., et al. (2008) Estimating individual-level exposure to airborne polycyclic aromatic hydrocarbons throughout the gestational period based on personal, indoor, and outdoor monitoring. Environ Health Perspect, 116: 1509–1518.
8. Crambes C, Kneip A, and Sarda P. (2009) Smoothing splines estimators for functional linear regression. Ann. Statist. Volume 37, Number 1: 35–72.
9. Davidian M, and Giltinan D. (1995) Nonlinear models for repeated measurement data. London: Chapman and Hall.
10. Elmi A, Ratcliffe S, Parry S, and Guo W. (2011) A B-Spline Based Semiparametric Nonlinear Mixed Effects Model. Journal of Computational and Graphical Statistics, 20(2): 492–509.
11. Fan J. and Zhang J.T. (2000) Two-Step Estimation of Functional Linear Models With Applications to Longitudinal Data. Journal of the Royal Statistical Society, Series B, 62: 303–322.
12. Fitzmaurice, G. M., Laird, N. M. and Ware, J. H. (2004). Applied Longitudinal Analysis. New Jersey: John Wiley & Sons.
13. Geman, A. and Hwang, C. (1982). Nonparametric maximum likelihood estimation by the method of sieves. Ann. Statist. 10: 401-414.
14. Guo, W. (2002). Functional Mixed Effects Models. Biometrics 58: 121–128.
15. Green, P. J. and Silverman B. W. (1994). Nonparametric Regression and Generalized Linear Models. London: Chapman and Hall.
16. James GM. (2002) Generalized linear models with functional predictors. Journal of the Royal Statistical Society, Series B, 64: 411–432.
17. Jedrychowski W, Whyatt RM, Camann DE, Bawle UV, Peki K, Spengler JD, et al. (2003) Effect of prenatal PAH exposure on birth outcomes and neurocognitive development in a cohort of newborns in Poland. Study design and preliminary ambient data. Int J Occup Med Environ Health, 16(1):21-29.
18. Jedrychowski W, Bendkowska I, Flak E, Penar A, Jacek R, Kaim I, et al. (2004) Estimated risk for altered fetal growth resulting from exposure to fine particles during pregnancy: an epidemiologic prospective cohort study in Poland. Environ Health Perspect, 112(14):1398–1402.
19. Jedrychowski WA, Perera FP, Pac A, Jacek R, Whyatt RM, Spengler JD, et al. (2006) Variability of total exposure to PM2.5 related to indoor and outdoor pollution sources. Krakow study in pregnant women. Sci Total Environ, 366(1): 47-54.
20. Junninen H, Mnster J, Rey M, Cancelinha J, Douglas K, et al. (2009) Quantifying the Impact of Residential Heating on the Urban Air Quality in a Typical European Coal Combustion Region. Environmental Science and Technology, 43: 7964–7970.
21. Ke, C., and Wang, Y. (2001) Semiparametric Nonlinear Mixed Effects Models and Their Applications (with discussion). Journal of the American Statistical Association, 96: 1272-1298.
22. Kim, J. J. (2004) Ambient air pollution: Health hazards to children. Pediatrics, 114: 1699–1707.

23. O 'Sullivan F, Yandall BS, and Raynor WJ. (1986) Automatic smoothing of regression functions in generalized linear models. J. Amer. Statist. Assoc. 81: 96–103.
24. Perera F, Tang D, Whyatt R, Lederman SA, and Jedrychowski W (2005) DNA damage from polycyclic aromatic hydrocarbons measured by benzo[a]pyrene-DNA adducts in mothers and newborns from Northern Manhattan, the World Trade Center Area, Poland, and China. Cancer Epidemiol Biomarkers Prev, 14: 709–714.
25. Ramsay JO, and Silverman BW. (1997) Functional Data Analysis. New York, Springer.
26. Ramsay JO, Hooker G, and Graves S. (2009) Functional Data Analysis with R and MATLAB. New York, Springer.
27. Sanyal MK, and Li YL. (2007) Deleterious effects of polynuclear aromatic hydrocarbon on blood vascular system of the rat fetus. Birth Defects Res B Dev Reprod Toxicol, 80: 367–373.
28. Schwartz, J. (2004) Air pollution and children's health. Pediatrics, 113: 1037–1043.
29. Selevan SG, Kimmel CA, and Mendola P. (2000) Identifying critical windows of exposure for children's health. Environ Health Perspect, 108: 451–455.
30. Sram, R. J., Binkova, B., Dejmek, J., and Bobak, M. (2005) Ambient Air Pollution and Pregnancy Outcomes: A Review of the Literature. Environmental Health Perspectives, 113: 375–382.
31. Shen, X. and Wong, W. (1994). Convergence rate of sieve estimates. Ann. Statist. 22: 580-615.
32. Stone, C. (1985). Additive regression and other nonparametric models. Ann. Statist. 13: 689-705.
33. Stone, C. J. (1986). The dimensionality reduction principle for generalized additive models. Ann. Statist. 14: 590–606.
34. Wahba G. (1990) Spline Models for Observational Data. SIAM, Philadelphia.
35. West LJ. (2002) Defining critical windows in the development of the human immune system. Hum Exp Toxicol, 21: no. 9-10: 499-505.
36. Yu Z, Loehr CV, Fischer KA, Louderback MA, Krueger SK, et al. (2006) In utero exposure of mice to dibenzo[a,l]pyrene produces lymphoma in the offspring: role of the aryl hydrocarbon receptor. Cancer Res, 66: 755–762.
37. Zhang, D., Lin, X. and Sowers, M. F. (2007) Two-Stage Functional Mixed Models for Evaluating the Effect of Longitudinal Covariate Profiles on a Scalar Outcome. Biometrics, 63: 351–362.
38. Zhang D., Lin X., Raz J., and Sowers M. (1998). Semiparametric stochastic mixed models for longitudinal data, Journal of the American Statistical Association, 93: 710–719.
39. Zhang, D., Lin, X. and Sowers, M. F. (2000). Semiparametric Regression for Periodic Longitudinal Hormone Data from Multiple Menstrual Cycles. Biometrics 56: 31–39.