Using semiparametric-mixed model and functional linear model to detect vulnerable prenatal window to carcinogenic polycyclic aromatic hydrocarbons on fetal growth

Lu Wang*,[1] and **Hyunok Choi**[2]

[1] Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA
[2] Department of Environmental Health Sciences, University at Albany, Rensselaer, NY 12144, USA

Received 10 June 2012; revised 14 July 2013; accepted 29 July 2013

Prenatal exposure to carcinogenic polycyclic aromatic hydrocarbons (c-PAHs) through maternal inhalation induces higher risk for a wide range of fetotoxic effects. However, the most health-relevant dose function from chronic gestational exposure remains unclear. 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, where a nonparametric cyclic smooth function plus a linear function are used to model the time effect and random effects are used to account for the within-subject correlation. We propose a penalized least squares approach to estimate the parametric regression coefficients and the nonparametric function of time. The smoothing parameter and variance components are selected using the generalized cross-validation (GCV) criteria. The estimated subject-specific trajectory of prenatal exposure is linked to the birth outcomes through a set of functional linear models, where the coefficient of log 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 ages, 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 function to draw statistical inferences, and the smoothing parameter is chosen through GCV.

Keywords: Environmental health; Longitudinal study; Risk assessment; Spline basis; Windows of vulnerability.

Additional supporting information may be found in the online version of this article at the publisher's web-site

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 (Barr et al., 2000; Selevan 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 pyrolysis of fossil fuel, coal, wood, cigarette, and food items, exerts both developmental toxicity, carcinogenicity and disruption of the endocrine system (Perera et al., 2005; Yu et al., 2006; Castro et al., 2008). Strong associations of prenatal exposure to PAHs with small-for-gestational age, preterm

*Corresponding author: e-mail: luwang@umich.edu, Phone: +1-734-647-6935, Fax: +1-734-763-2215 delivery, and neuro-developmental deficits in children have also been observed (Selevan et al., 2000; Kim, 2004; Schwartz, 2004). 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, the 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 latter half of the pregnancy period has been deemed the most detrimental. However, a 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 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 the 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 units of exposure across different gestational ages.

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-hour 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 to provide a data-driven 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). We use a cyclic nonparametric smooth function plus a linear function of the calendar

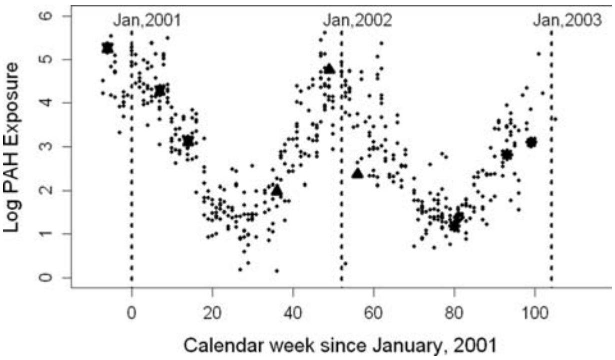

Figure 1 Personal prenatal log PAH exposure measurements across the calendar time among the study cohort from November 2000 to January 2003. The observations of four subjects are illustrated using different subject-specific symbols.

time to model the longitudinal trajectory of log PAH exposure, and use other parametric components to model the effect of other covariates. The within-subject correlation is accounted by subject-specific random effects. We employ penalized least squares methods to estimate the regression coefficients and the cyclic nonparametric function, by generalizing the method in Gu and Ma (2005) for nonparametric mixed-effect models. 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 log PAH exposure in functional linear models is a fully nonparametric function of gestational age, which allows the effect of log PAH exposure on each birth outcome to vary at different gestational ages, 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 parameters are chosen through the generalized cross-validation (GCV) criteria.

2 Statistical models

Eight carcinogenic PAHs were monitored and measured during the unique 48-hour 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 periodic effect of calendar time on the PAH exposure level as shown in Fig. 1, we fit the following semiparametric-mixed effect model of log PAH with a subject-specific random intercept and random slope for some covariates.

$$Y_{ij} = X_{ij}^T \boldsymbol{\beta} + f(t_{ij}) + b_{i1} + Z_{ij}^T \boldsymbol{b}_2 + e_{ij}, \tag{1}$$

where Y_{ij} denotes the jth log PAH exposure measurement for subject i at time t_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, n_i$, 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 X_{ij}. We model the effect of calender time t_{ij} on the log PAH exposure level through a nonparametric cyclic function $f(t_{ij})$ to allow for any possible periodic nonlinear association, where $f(t)$ is assumed to be a twice-differentiable periodic smooth function. Besides the cyclic pattern captured by $f(t)$, we also allow a linear change of log PAH exposure over calendar time. Thus, t_{ij} is included in X_{ij} as well. In order to make the model identifiable, we restrict $f(t)$ to have mean 0. The effect of other potential predictors, such as spatial factors (whether living in city center) and behavioral factors (whether smoking), are also modeled linearly through $X_{ij}^T\boldsymbol{\beta}$ in the model. We use the subject specific random intercept b_{i1}, $i = 1, \ldots, n$, and $\boldsymbol{b}_2$, the random effects corresponding to Z_{ij}, to handle the within-subject correlation. In our analysis, Z_{ij} includes four dummy variables to indicate whether living in city center and whether smoking. Thus, $\boldsymbol{b}_2 = (b_{21}, b_{22}, b_{31}, b_{32})^T$. Denote $\boldsymbol{b} = (b_{11}, b_{21}, \ldots, b_{n1}, b_{21}, b_{22}, b_{31}, b_{32})^T$, and we assume $\boldsymbol{b} \sim N\{\boldsymbol{0}, B(\boldsymbol{\theta})\}$, where $\boldsymbol{\theta}$ denotes the variance components. $e_{ij} \sim N(0, \sigma_e^2)$ are independent measurement errors. Using such a semiparametric-mixed effect model, each subject borrows information from the other subjects in the study cohort to fit the whole personal PAH exposure profile. Note that the covariates in Z_{ij} could be time dependent. Thus, 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 departure as $Z_{ij}^T\boldsymbol{b}_2$ corresponding to the value of Z_{ij} for subject i at time t_{ij}.

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

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

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 log PAH exposure profile predicted from model (1), and $\alpha(t)$ is the coefficient function associated with log 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 at delivery T_d. $\alpha(t)$ is a fully nonparametric function of t, which allows the effect of log PAH exposure on each birth outcome to vary at different gestational ages. We controlled for other potential risk factors such as the gestational age, newborn gender, parity, whether the delivery is c-section, whether born in summer season (from April to September), mom's prepregnancy weight as well as maternal height through $\widetilde{\boldsymbol{X}}_i^T\boldsymbol{\gamma}$, where $\widetilde{\boldsymbol{X}}_i$ denotes 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 $\widetilde{\boldsymbol{X}}_i$. We assume $\epsilon_i \sim N(0, \sigma_\epsilon^2)$ are independent random errors.

3 Estimation procedure

3.1 Penalized least squares estimation in semiparametric-mixed effect model

To facilitate the presentation, we introduce the following matrix notation. Let $\boldsymbol{Y}_i = \left(Y_{i1}, \ldots, Y_{in_i}\right)^T$, for $i = 1, \cdots n$, and similarly define $\boldsymbol{X}_i$, $\boldsymbol{Z}_i$, $\boldsymbol{t}_i$, and $\boldsymbol{e}_i$. Considering all the subjects in the study and further denoting $Y = \left(\boldsymbol{Y}_1^T, \ldots, \boldsymbol{Y}_n^T\right)^T$, $X = \left(\boldsymbol{X}_1^T, \ldots, \boldsymbol{X}_n^T\right)^T$, $t = \left(\boldsymbol{t}_1^T, \ldots, \boldsymbol{t}_n^T\right)^T$, $Z = [diag(1_{n_1}, \ldots, 1_{n_n}), (\boldsymbol{Z}_1^T, \ldots, \boldsymbol{Z}_n^T)^T]$, $\boldsymbol{b} = (b_{11}, b_{21}, \ldots, b_{n1}, b_{21}, b_{22}, b_{31}, b_{32})^T$, and $e = \left(\boldsymbol{e}_1^T, \ldots, \boldsymbol{e}_n^T\right)^T$, we have

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

where $\boldsymbol{\beta}$ and f are defined as before, $\boldsymbol{b} \sim N\{\boldsymbol{0}, B(\boldsymbol{\theta})\}$, $\boldsymbol{\theta}$ denotes the variance components for all random effects, $e \sim N\left\{0, \sigma_e^2 I_N\right\}$, and I_N is the identity matrix of dimension $N = \sum_{i=1}^n n_i$. We consider an expansion of $f(t)$ in a r-dimensional space

$$f(t) = \boldsymbol{\xi}(t)^T\boldsymbol{c}$$

where $\boldsymbol{\xi}(t) = \{\xi_1(t), \cdots \xi_r(t)\}$ are basis functions, and $\boldsymbol{c} = (c_1, \ldots, c_r)^T$ are coefficients for the linear expansion. In our case, since $f(t)$ is assumed to be twice differentiable and cyclic with the cycle as one year in model (1), we choose to use the cyclic cubic regression spline basis. Generalizing the estimating procedure of Gu and Ma (2005) for nonparametric mixed effect models to include a parametric component, we minimize the following penalized least squares with respect to $\boldsymbol{\beta}$, $\boldsymbol{b}$, and $\boldsymbol{c}$.

$$L(\boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{c}) = (Y - X\boldsymbol{\beta} - Z\boldsymbol{b} - M\boldsymbol{c})^T\,(Y - X\boldsymbol{\beta} - Z\boldsymbol{b} - M\boldsymbol{c}) + \sigma_e^2 \boldsymbol{b}^T B(\boldsymbol{\theta})^{-1}\boldsymbol{b} + N\lambda \boldsymbol{c}^T Q\boldsymbol{c}, \qquad (4)$$

where $M = [\boldsymbol{\xi}(t_{11})^T, \cdots \boldsymbol{\xi}(t_{nn_n})^T]^T$, λ is the smoothing parameter, a nonnegative constant, and $\boldsymbol{Q}$ is the penalty matrix corresponding to the roughness of the fitted curve. We treat the variance components $\boldsymbol{\theta}$ and σ_e^2 together with λ as tuning parameters, which will be determined through generalized cross-validation. We choose to use roughness penalty as the total curvature of $f(t)$. Specifically, the (i, j)th element of Q is

$$Q_{ij} = \int_{T_1}^{T_2} \xi_i''(t)\xi_j''(t)dt,$$

where the integrals are taken over the domain of spline basis (T_1, T_2), and $\xi_k''(t)$ is the second derivative of basis function $\xi_k(t)$. Differentiating (4) with respect to $\boldsymbol{\beta}$, $\boldsymbol{b}$, and $\boldsymbol{c}$, and setting the derivatives to 0, we solve

$$\begin{bmatrix} X^T X & X^T Z & X^T M \\ Z^T X & Z^T Z + \sigma_e^2 B(\boldsymbol{\theta})^{-1} & Z^T M \\ M^T X & M^T Z & M^T M + \lambda N Q \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{b} \\ \boldsymbol{c} \end{bmatrix} = \begin{bmatrix} X^T Y \\ Z^T Y \\ M^T Y \end{bmatrix}$$

to obtain the estimators $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{b}}, \widehat{\boldsymbol{c}})$. Then $X_{i,t}^T\widehat{\boldsymbol{\beta}} + \boldsymbol{\xi}(t)^T\widehat{\boldsymbol{c}} + Z_{i,t}^T\widehat{\boldsymbol{b}}$ is the estimated longitudinal trajectory of personal log PAH exposure for subject i, where $X_{i,t}$ and $Z_{i,t}$ are the most recent values for the corresponding covariates at time t. Note

$$\widehat{Y} = \begin{pmatrix} X & Z & M \end{pmatrix} \begin{pmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{b}} \\ \widehat{\boldsymbol{c}} \end{pmatrix} = A(\theta, \sigma_e^2, \lambda)Y$$

where

$$A(\boldsymbol{\theta}, \sigma_e^2, \lambda) = \begin{pmatrix} X & Z & M \end{pmatrix} \begin{bmatrix} X^T X & X^T Z & X^T M \\ Z^T X & Z^T Z + \sigma_e^2 B(\boldsymbol{\theta})^{-1} & Z^T M \\ M^T X & M^T Z & M^T M + \lambda N Q \end{bmatrix}^{-1} \begin{bmatrix} X^T \\ Z^T \\ M^T \end{bmatrix}$$

is a function of $\boldsymbol{\theta}, \sigma_e^2,$ and λ. We choose the optimal λ and simultaneously estimate the variance components $(\boldsymbol{\theta}, \sigma_e^2)$ by minimizing the generalized cross-validation (GCV) criteria (Craven and Wahba, 1979)

$$V(\boldsymbol{\theta}, \sigma_e^2, \lambda) = \frac{N^{-1} Y^T \left[I - A(\boldsymbol{\theta}, \sigma_e^2, \lambda)\right]^2 Y}{\left\{N^{-1} tr\left[I - A(\boldsymbol{\theta}, \sigma_e^2, \lambda)\right]\right\}^2},$$

which, as shown in Gu and Ma (2005), yields the optimal smoothing. Using such GCV criteria also greatly reduces the computational intensity in practice (Ramsay et al., 2009, among others). The optimal fitting depends on r, the dimension of basis functions, as well, and we choose r to globally minimize the GCV score.

Zhang et al. (2007) proposed to convert the "roughness" term to a random effect, and hence treat the minimization problem as a linear mixed model. However, our experience is that the numerical performance of this method is not very stable. We employed the GCV criteria as our objective function to select $\boldsymbol{\theta}, \sigma_e^2,$ and λ. Note that $V(\boldsymbol{\theta}, \sigma_e^2, \lambda)$ is flat with respect to very small and very large smoothing parameters, and thus the general optimization algorithm may stuck in a local area. One can further assume some properties for $B(\boldsymbol{\theta})$ to facilitate the optimization implementation. We choose to use simple random effects, and assume that $B(\boldsymbol{\theta})$ is a block diagonal matrix of the form

$$\begin{bmatrix} \theta_1 I_1 & 0 & \cdots & 0 \\ 0 & \theta_2 I_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \theta_k I_k \end{bmatrix},$$

where k is the dimension of $\boldsymbol{\theta}$. The penalties thus become a series of quadratic terms with known penalty matrixes and unknown tuning parameters. We then take advantage of the function “magic” in R package, mgcv. This function utilizes Newton’s method in multi-dimensions, combined with steepest descent to iteratively update the smoothing parameters for each penalty. We found that this approach has superior numerical stability.

3.2 Estimation in functional linear models using spline basis

Functional linear model (Ramsay and Silverman, 1997; Cardot, 2003; Ramsay et al., 2009) 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 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 (Wahba, 1990; Ramsay and Silverman, 1997), using a similar expansion of $\alpha(t)$ but with fixed number of basis functions. Similarly as in Section 3.1, 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 that has continuous second-order derivative. We minimize the following penalized sum of squared errors

$$\sum_{i=1}^{n} \left\{ O_i - \int_0^{T_d} \widehat{Y}_i(t)\alpha(t)dt - \widetilde{\boldsymbol{X}}_i^T \boldsymbol{\gamma} \right\}^2 + \rho \int_0^{T_d} \left\{\alpha''(t)\right\}^2 dt, \qquad (5)$$

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 spline basis $\left\{S_k(t),\ k = 1, \ldots, K_\alpha\right\}$, where K_α is the number of basis functions. We represent $\alpha(t)$ as

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

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

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

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

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

Thus, the penalized residual sum of squares (5) can be rewritten as

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

where $S_k''(t)$ is the second derivative of spline function $S_k(t)$. Let $\boldsymbol{U}_{K_\alpha}(t) = \left\{S_1''(t), \ldots, S_{K_\alpha}''(t)\right\}^T$ and let 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 $\rho \boldsymbol{\phi}^T R \boldsymbol{\phi}$, similarly as in Section 3.1. We can further simplify notation by defining $\zeta = \left(\boldsymbol{\phi}^T, \boldsymbol{\gamma}^T\right)^T$, $W_i = \left(\boldsymbol{J}_i^T, \widetilde{\boldsymbol{X}}_i^T\right)^T$, and R_0 as a $\left(K_\alpha + d_2\right) \times \left(K_\alpha + d_2\right)$ matrix, which augments R by attaching **0**s. Then the penalized objective function (5) that we want to minimize becomes

$$\sum_{i=1}^{n} \left\{O_i - W_i^T \zeta\right\}^2 + \rho \zeta^T R_0 \zeta.$$

We can estimate ζ by solving

$$\left(W^T W + \rho R_0\right) \zeta = W^T \boldsymbol{O},$$

where $\boldsymbol{O} = \left(O_1, \ldots, O_n\right)^T$ is a $n \times 1$ vector, and $W = \left(W_1, \ldots, W_n\right)^T$ is a $n \times \left(K_\alpha + d_2\right)$ matrix. Correspondingly, we can construct confidence intervals for $\widehat{\zeta}$. We estimate the variance of $\widehat{\zeta}$ as

$$\widehat{Var}\left(\widehat{\zeta}\right) = \widehat{\sigma}_\epsilon^2 \left(W^T W + \rho R_0\right)^{-1} W^T W \left(W^T W + \rho R_0\right)^{-1},$$

and $\widehat{\sigma}_\epsilon^2$ can be calculated from the mean squared residuals. The algorithm described in Section 3.1 can be applied here as well with slight modifications. The smoothing parameter ρ and the number of basis functions K_α are chosen similarly as in Section 3.1 for λ and r by the GCV criteria.

4 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 ($n = 344$) were simultaneously monitored for their personal, 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

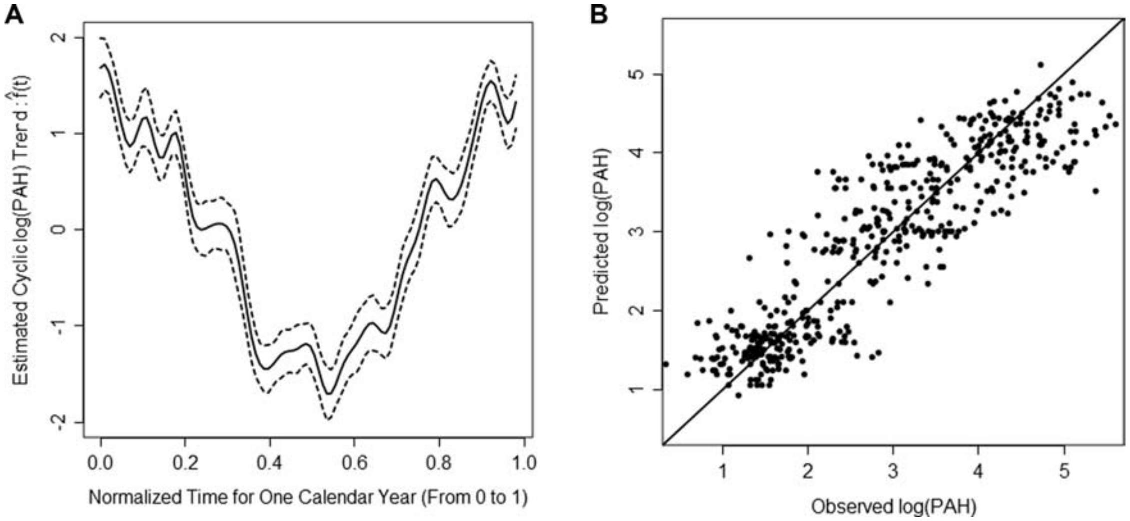

Figure 2 Estimation of the cyclic nonparametric function $f(t)$ in one cycle (— is the estimate, and ---- is the 95% confidence interval) and the predictions of individual prenatal PAH exposure during pregnancy based on semiparametric-mixed effect model. (A) Estimation of cyclic function $f(t)$ in one cycle (b) predicted versus observed PAH.

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 that operated for a consecutive 48-hour period. The split flow inlet, placed near the woman's breathing zone, drew in the particulate or semivolatile vapor PAHs and particles 2.5 m (PM2.5) on a precleaned 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).

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 from different trimesters. 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 symbols, 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 Fig. 1 demonstrating a nonlinear periodic effect of calendar time on log PAH exposure, which is reasonable given that the study was conducted for over 2 years. Due to the heating mechanism in Krakow, the PAH exposure level is usually higher in winter and lower in summer, and thus demonstrate a cyclic pattern over the study period. This observable nonlinear periodic effect of the calendar time on the log PAH exposure level is captured by the semiparametric model (1) of log PAH with a cyclic function $f(t)$. From our experience, the level of wiggliness of $\widehat{f}(t)$ directly connects to the number of knots used when constructing the spline basis, and the smoothing parameter λ. We apply the GCV criteria (smaller is better) and perform a comprehensive search on the number of knots, from 4 to 50. For each fixed number of knots, we choose the optimal λ by minimizing GCV. Then we choose the number of knots that achieves the lowest GCV score. The estimated function $\widehat{f}(t)$ in one cycle, together with the 95% confidence interval, is displayed in Fig. 2A.

Besides the cyclic effects, our results also show evidence for a decreasing trend of the log PAH exposure over calendar time (p-value$<$ 0.01). This demonstrates that the air pollution in this area has been improved over the years. 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 0.068 units in log PAH scale (p-value= 0.94) while smoking increases the risk of being exposed to higher PAH level by 0.061 units in log PAH scale (p-value$<$ 0.01). These results suggest that, in addition to the seasonal effect and calendar time effect, host's personal behavior also critically influences the magnitude of exposure risk to PAH. We compare all observed measurements of log PAH exposure in the study versus the predicted values in Fig. 2B, and most points fall along the diagonal line nicely.

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 log PAH exposure trajectory. This can be viewed as a step of curve registration, where we align the estimated exposure profiles according to gestational age. We preserve the first two trimesters in order to have a biologically meaningful interpretation and only rescale the last trimester so that the length of the normalized prenatal exposure period is the same for everyone. This makes the integral limit in model (2) nonrandom 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 3 displays the estimated regression coefficient function, $\widehat{\alpha}(t)$, along with 95% point-wise confidence intervals. For each birth outcome, the risk of intrauterine growth restriction due to prenatal PAH exposure is a function of gestational period. Negative values of $\widehat{\alpha}(t)$ means adverse effect, and the magnitude of $\widehat{\alpha}(t)$ is the loss of baby's birth outcome associated with one unit increase of prenatal log PAH exposure at gestational age t. The lower $\widehat{\alpha}(t)$ is, the higher the risk associated with PAH exposure is at that specific gestational age t. For log 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 growth and length gain. However, birth head circumference is affected more and more detrimentally across the gestational age by the occurrence of PAH exposure, which is not surprising given that brain and neural system start to develop rapidly in the second trimester. For fetal weight development, a significant window of vulnerability is identified from around gestational week 3 to week 18, where the confidence interval does not contain zero in Fig. 3A. From the same plot, one can see that the adverse effect of PAH exposure on birth weight is also significant in part of the third trimester, from around week 30 to delivery in the standardized gestational age scale. For fetal length growth, Fig. 3B shows that the adverse effect of PAH exposure is statistically significant in both the first trimester (from fertilization to around gestational week 14) and in the third trimester (from around week 27 to delivery). No significant window of vulnerability was shown from this study for fetal head circumference development (Fig. 3C). We controlled for the other potential risk factors including whether the baby was born in summer, the gestational age at born, newborn gender, parity, if the delivery is c-section or not, prepregnancy weight, as well as maternal height. The results in Table 1 show that baby boys tend to have a significantly higher birth weight (p<0.001), longer birth length (p<0.001), and larger birth head circumference (p<0.001) than baby girls. Heavier prepregnancy weight of mom, higher maternal height, and longer gestational period are significantly associated with an increase of birth weight, birth length, and birth head circumference (all p-values <0.05). Moms who have had baby before tend to bear a baby with higher weight and larger head circumference (p-value are 0.038 and 0.001, respectively). If the delivery is a c-section, the newborn tends to have a larger birth head circumference (p-value = 0.027). We also controlled c-section delivery in the model of birth weight and birth length, but no statistical significance is observed. Whether the baby was born in the winter season (from October to March, when heating becomes necessary and nutrition might become different) is negatively correlated with all birth outcomes, but none of the associations is statistically significant.

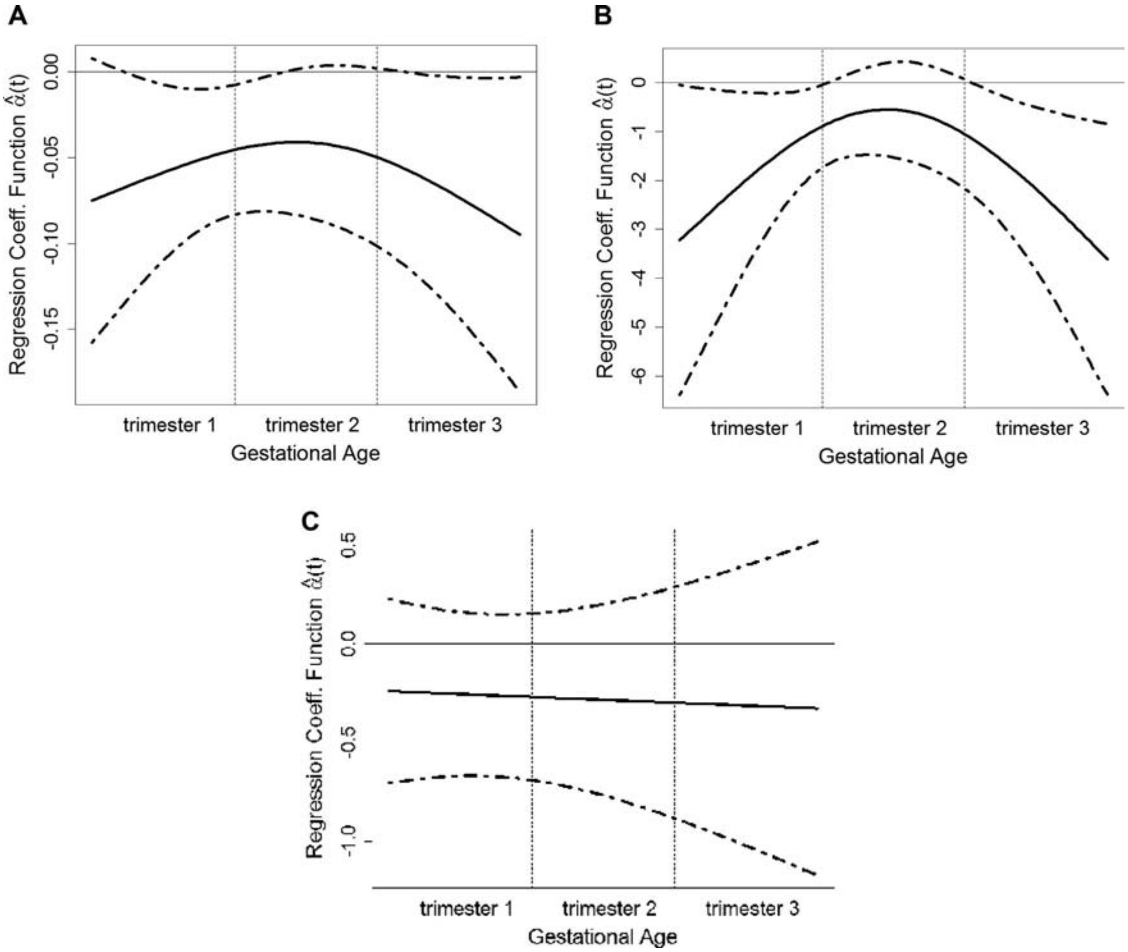

Figure 3 The estimated regression coefficient function $\alpha(t)$ in functional linear model, that is, the effect of prenatal log PAH exposure on baby's log birth weight, birth length, and birth head circumference across gestational age t. —— is the estimate and – · – · – is the 95% confidence interval. (A) log Birth Weight, (B) Birth Length, (C) Birth Head Circumference.

Table 1 Estimates of the effects of other risk factors in the functional linear models of birth outcomes.

	log(Birth Weight)		Birth Length		Birth H-C	
	coefficient	*p*-value	coefficient	*p*-value	coefficient	*p*-value
Maternal height	0.002	0.041	0.056	0.025	0.030	0.023
Prepregnancy weight	0.003	<0.001	0.046	0.003	0.027	<0.001
log(Gestational Age)	2.261	<0.001	36.37	<0.001	14.10	<0.001
Parity (yes versus no)	0.027	0.038	0.364	0.188	0.474	0.001
Newborn gender (girl versus boy)	−0.058	<0.001	−1.078	<0.001	−0.733	<0.001
Whether c-section delivery	−0.020	0.204	0.278	0.421	0.403	0.027
Whether born in summer season	0.029	0.224	0.001	0.993	0.120	0.639

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 so 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 ages 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 with a nonparametric cyclic function together with a linear function of calendar time, and random effects are used to account for within-subject correlations. Using curve registration, the estimated subject specific trajectory of prenatal log PAH exposure 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 ages. The window associated with significantly adverse effects is identified as a vulnerable prenatal window to PAHs on fetal growth. To draw statistical inferences in both longitudinal semiparametric mixed effect model and functional linear models, we minimize the penalized least squares objective function using a spline-based expansion of the nonparametric functions. The smoothing parameters are selected using GCV criteria.

Our results show that prenatal PAH exposure is associated with reduction in birth weight, birth length, as well as birth head circumference. There is evidence in this study suggesting that the vulnerability of the fetus against high prenatal PAH exposure varies across different gestational ages, and thus one may want to avoid unnecessary toxic exposures accordingly to reduce the potential risk. Specifically, our results hint at a couple of critical windows of vulnerability for fetal weight and height development, during which the PAH exposure yields significant impairment, and thus reducing PAH exposure during these gestational windows may help for fetal weight and length development. For birth head circumference, it appears to be affected more and more detrimentally across the gestational age, but no statistical significance is found.

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 and the third 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. One limitation of this study design is the fact that the repeated measurements in our data are very sparse due to technical difficulties related to pregnancy burden and costly expense considerations. Therefore, the advantage of using random effects to account for within-subject correlations is limited, even though we notice a slight decrease of the GCV score for the model that includes the random effects (e.g. GCV-score = 0.363 for the model with random intercept, while in the model without random intercept, GCV-score = 0.364). However, we believe that for other studies with similar goals and well designed with more repeated measurements, our proposed models and methods will provide more significant insights.

Acknowledgments The authors wish to thank the editor, the associate editor, and two reviewers for their very helpful comments.

Conflict of interest

The authors have declared no conflict of interest.

References

Barr, M., DeSesso, J. M., Lau, C. S., Osmond, C., Ozanne, S. E., Sadler, T. W., Simmons, R. A. and Sonawane, B. R. (2000). Workshop to identify critical windows of exposure for children's health: cardiovascular and endocrine work group summary. *Environmental Health Perspectives* **108**, 569–571.

Cardot, H., Ferraty, F. and Sarda, P. (1999). Functional linear model. *Statistics & Probability Letters* **45**, 11–22.

Cardot, H., Ferraty, F. and Sarda, P. (2003). Spline estimators for the functional linear model. *Statistica Sinica* **13**, 571–591.

Cardot, H. and Sarda, P. (2005). Estimation in generalized linear models for functional data via penalized likelihood. *Journal of Multivariate Analysis* **92**, 24–41.

Castro, D. J., Lohr, C. V., Fischer, K. A., Pereira, C. B. and Williams, D. E. (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. *Toxicology and Applied Pharmacology* **233**, 454–458.

Choi, H., Jedrychowski, W., Spengler, J., Camann, D. E., Whyatt, R. M., Rauh, V., Tsai, W. and Perera, F. P. (2006). International studies of prenatal exposure to polycyclic aromatic hydrocarbons and fetal growth. *Environmental Health Perspectives* **114**, 1744–1750.

Choi, H., Perera, F., Pac, A., Wang, L., Flak, E., Mroz, E., Jacek, R., Chai-Onn, T., Jedrychowski, W., Masters, M., Camann, D. and Spengler, J. (2008). Estimating individual-level exposure to airborne polycyclic aromatic hydrocarbons throughout the gestational period based on personal, indoor, and outdoor monitoring. *Environmental Health Perspectives* **116**, 1509–1518.

Crambes, C., Kneip, A. and Sarda, P. (2009). Smoothing splines estimators for functional linear regression. *Annals of Statistics* **37**, 35–72.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross- validation. *Numerische Mathematik* **31**, 377–403.

Davidian, M. and Giltinan, D. (1995). *Nonlinear Models for Repeated Measurement Data*. Chapman and Hall, London, UK.

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**, 492–509.

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.

Fitzmaurice, G. M., Laird, N. M. and Ware, J. H. (2004). *Applied Longitudinal Analysis*. John Wiley & Sons, New Jersey, NJ.

Geman, A. and Hwang, C. (1982). Nonparametric maximum likelihood estimation by the method of sieves. *Annals of Statistics* **10**, 401–414.

Gu, C. and Ma, P. (2005). Optimal smoothing in nonparametric mixed-effect models. *Annals of Statistics* **33**, 1357–1379.

Guo, W. (2002). Functional mixed effects models. *Biometrics* **58**, 121–128.

Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models*. Chapman and Hall, London, UK.

James, G. M. (2002). Generalized linear models with functional predictors. *Journal of the Royal Statistical Society, Series B* **64**, 411–432.

Jedrychowski, W., Whyatt, R. M., Camann, D., Bawle, U. V., Peki, K., Spengler, J. D., Dumyahn, T., Penar, A. and Perera, F. (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. *International Journal of Occupational Medicine and Environmental Health* **16**, 21–29.

Jedrychowski, W., Bendkowska, I., Flak, E., Penar, A., Jacek, R., Kaim, I., Spengler, J. D., Camann, D. and Perera, F. (2004). Estimated risk for altered fetal growth resulting from exposure to fine particles during pregnancy: an epidemiologic prospective cohort study in Poland. *Environmental Health Perspectives* **112**, 1398–1402.

Jedrychowski, W. A., Perera, F. P., Pac, A., Jacek, R., Whyatt, R. M., Spengler, J. D., Dumyahn, T. and Sochacka-Tatara, E. (2006). Variability of total exposure to PM2.5 related to indoor and outdoor pollution sources. Krakow study in pregnant women. *Science of The Total Environment* **366**, 47–54.

Junninen, H., Monster, J., Rey, M., Cancelinha, J., Douglas, K., Duane, M., Forcina, V., Mller, A., Lagler, F., Marelli, L., Borowiak, A., Niedzialek, J., Paradiz, B., Mira-Salama, D., Jimenez, J., Hansen, U., Astorga, C., Stanczyk, K., Viana, M., Querol, X., Duvall, R. M., Norris, G. A., Tsakovski, S., Whlin, P., Hork, J., Larsen, B. R. (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.

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.

Kim, J. J. (2004). Ambient air pollution: health hazards to children. *Pediatrics* **114**, 1699–1707.

O 'Sullivan, F., Yandall, B. S. and Raynor, W. J. (1986). Automatic smoothing of regression functions in generalized linear models. *Journal of the American Statistical Association* **81**, 96–103.

Perera, F., Tang, D., Whyatt, R., Lederman, S. A. 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 Epidemiology, Biomarkers, & Prevention* **14**, 709–714.

Ramsay, J. O. and Silverman, B. W. (1997). *Functional Data Analysis*. Springer, New York, NY.

Ramsay, J. O., Hooker, G. and Graves, S. (2009). *Functional Data Analysis with R and MATLAB*. Springer, New York, NY.

Sanyal, M. K. and Li, Y. L. (2007). Deleterious effects of polynuclear aromatic hydrocarbon on blood vascular system of the rat fetus. *Birth Defects Research Part B: Developmental and Reproductive Toxicology* **80**, 367–373.

Schwartz, J. (2004). Air pollution and children's health. *Pediatrics* **113**, 1037–1043.

Selevan, S. G., Kimmel, C. A. and Mendola, P. (2000). Identifying critical windows of exposure for children's health. *Environmental Health Perspectives* **108**, 451–455.

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.

Shen, X. and Wong, W. (1994). Convergence rate of sieve estimates. *Annals of Statistics* **22**, 580–615.

Stone, C. J. (1985). Additive regression and other nonparametric models. *Annals of Statistics* **13**, 689–705.

Stone, C. J. (1986). The dimensionality reduction principle for generalized additive models. *Annals of Statistics* **14**, 590–606.

Wahba, G. (1990). *Spline Models for Observational Data*. SIAM, Philadelphia, PA.

West, L. J. (2002). Defining critical windows in the development of the human immune system. *Human & Experimental Toxicology* **21**, 499–505.

Yu, Z., Loehr, C. V., Fischer, K. A., Louderback, M. A., Krueger, S. K., Dashwood, R. H., Kerkvliet, N. I., Pereira, C. B., Jennings-Gee, J. E., Dance, S. T., Miller, M. S., Bailey, G. S. and Williams, D. E. (2006). In utero exposure of mice to dibenzo[a,l]pyrene produces lymphoma in the offspring: role of the aryl hydrocarbon receptor. *Cancer Research* **66**, 755–762.

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.

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.

Zhang, D., Lin, X. and Sowers, M. F. (2000). Semiparametric regression for periodic longitudinal hormone data from multiple menstrual cycles. *Biometrics* **56**, 31–39.