Quantifying Risk Over the Life Course – Latency, Age-Related Susceptibility, and Other Time-Varying Exposure Metrics

HHS Public Access

Author manuscript

Stat Med. Author manuscript; available in PMC 2017 June 15.

Published in final edited form as:

Stat Med. 2016 June 15; 35(13): 2283–2295. doi:10.1002/sim.6864.

Quantifying Risk Over the Life Course – Latency, Age-Related Susceptibility, and Other Time-Varying Exposure Metrics

Molin Wang[1,*], **Xiaomei Liao**[2], **Francine Laden**[3], and **Donna Spiegelman**[4]

[1,2,4]Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA

[1,3,4]Department of Epidemiology, Harvard School of Public Health, Boston, MA, USA

[3]Department of Environmental Health, Harvard School of Public Health, Boston, MA, USA

[4]Department of Nutrition, Harvard School of Public Health, Boston, MA, USA

[1,3]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA

Abstract

Identification of the latency period and age-related susceptibility, if any, is an important aspect of assessing risks of environmental, nutritional and occupational exposures. We consider estimation and inference for latency and age-related susceptibility in relative risk and excess risk models. We focus on likelihood-based methods for point and interval estimation of the latency period and age-related windows of susceptibility coupled with several commonly considered exposure metrics. The method is illustrated in a study of the timing of the effects of constituents of air pollution on mortality in the Nurses' Health Study.

Keywords

latency; time to event data; cohort studies; time-varying exposure; Cox proportional hazard model

1. Introduction

Epidemiologists are often interested in estimating the effect of time-varying exposure variables in relation to disease endpoints, such as cancer and cardiovascular disease incidence and mortality. An exposure-disease relationship may be modified by temporal factors. Take for example, an instantaneous exposure, such as radiation dose resulting from an atomic bomb explosion [1, 2]. As discussed in Chapter 6 of [3], temporal factors that could modify this exposure-disease relationship include age at exposure, attained age, time since exposure, and calendar year of birth or risk. For time-varying exposures, with the availability of long-term exposure histories, temporal modifiers of the exposure effects can also be identified. For example, life course epidemiologists have investigated the association between early–, mid– and later life exposures and the risk of many diseases in later life, such

[*]Correspondence to: Departments of Biostatistics and Epidemiology, Harvard School of Public Health, and Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. stmow@channing.harvard.edu.

as type 2 diabetes and breast cancer [4]. In many settings, the timing of exposure may be a major modifier; for example, there may be a critical time window for an exposure, during which the risk of developing disease depends, rather than the risk varying uniformly with the exposure level over the entire life course. Identification of the beginning and end of this critical period of susceptibility, if any, is an important aspect of a comprehensive assessment of the public health effects of an environmental, nutritional or occupational exposure. Our motivating example arises from a study of the relationship between fine particulate matter < 2.5 μm in diameter ($PM_{2.5}$) and all-cause mortality in the Nurses Health Study (NHS), an ongoing prospective cohort of 121,700 US nurses who have been followed biannually since 1976 [5]. Here, interest is in estimating the critical time window of susceptibility, and the effect of $PM_{2.5}$ on all-cause mortality during this critical time window of susceptibility. Previous data have suggested that the exposure has an effect which begins a few years preceding the present time and ends at the present [6, 5, 7]. Another useful type of exposure metric defines the time window of susceptibility to begin at a certain age and end at the present time; this exposure metric may be used, for example, in the study of the body size in relation to breast cancer risk [8]. A further option allows the time window of susceptibility to end a few years preceding the present time to allow for a lag, where the most recent exposures are excluded due to the unlikely consideration of acute effects, as would be the case for many cancers. See more discussions of the exposure metrics in Section 2.

The strongest effect method has been used for estimating the latency period, which is the interval between the beginning of the exposure to development of the disease. In this method, the point estimate of the latency period is the one corresponding to the largest relative risk [9]. This method rests on the argument that the maximal effect estimate will be the one that is least biased by non-differential exposure misclassification. As discussed in [10] and [11], which focus on lag intervals, the strongest effect method will produce biased estimates, and the authors of both papers proposed a likelihood-based goodness-of-fit method to estimate the lag period. Salvan and others [10] considered the analysis of unmatched case-control studies with an unconditional logistic regression model and a binary exposure, and Richardson and others [11] considered the analysis of nested or matched case-control studies with a linear excess rate ratio model.

Let $c(t)$ be the value at time t of a time-varying exposure of interest, and $\mathbf{U}(t)$ be a column vector of potential confounders. When studying the timing of exposure, e.g., a latency period or age-related susceptibility, an appropriate exposure metric needs to be specified. Denote the pre-specified exposure metric as $X(\mathbf{c}(t); \mathbf{a})$, where $\mathbf{c}(t)$ is the history of the time-varying exposure levels observed up to time t, and $\mathbf{a}$ is a possibly vector-valued unknown parameter. Two forms of hazard rate models are

$$\text{relative risk model:} \lambda(t) = \lambda_0(t) r\{\beta X(\mathbf{c}(t); \mathbf{a}) + \boldsymbol{\beta}_u \mathbf{U}(t)\}, \quad (1)$$

$$\text{excess risk model:} \lambda(t) = \lambda_0(t) + d\{\beta X(\mathbf{c}(t); \mathbf{a}) + \boldsymbol{\beta}_u \mathbf{U}(t)\}, \quad (2)$$

where $\lambda(t)$ is the incidence rate at time t, $\lambda_0(t)$ is the baseline incidence rate at t, $r(\cdot)$ and $d(\cdot)$ are any real-valued functions, and β and $\boldsymbol{\beta}_u$ are unknown parameters. The models above are

subject to the constraint that $\lambda(t) > 0$ for all possible values of t and the model covariates. For presentational simplicity, we do not consider an interaction between $X(\mathbf{c}(t); \mathbf{a})$ and $\mathbf{U}(t)$. The methods described in this paper can easily be extended to the cases when there are $X(\mathbf{c}(t); \mathbf{a}) \times \mathbf{U}(t)$ interactions by treating the interaction term as a new time-varying variable.

For example, the Cox model, commonly used in the analysis of cohort studies, is a special case of the relative risk model. A standard Cox model with exposure metric $X(\mathbf{c}(t); \mathbf{a})$ is

$$\lambda(t)=\lambda_0(t)\exp\{\beta X(\mathbf{c}(t);\mathbf{a})+\boldsymbol{\beta}_u\mathbf{U}(t)\}. \quad (3)$$

where β and $\boldsymbol{\beta}_u$, a row vector, are the log relative risks (RR) (hazard or incidence rate ratios). When applying the Cox model in epidemiologic cohort studies of chronic disease, t is typically age as recommended by a number of authors [12, 13], and the model will be left-truncated [14]. Note that the time scale in the Cox regression model is not necessarily the same as the scale used to assess susceptibility or latency. When the two time scales are different, the time scale for assessing susceptibility and latency can be typically transformed to that used in the Cox model. For example, assuming the time scale in the Cox model is age, denoted as t, and the scale for assessing susceptibility and latency is time since the beginning of exposure, the latter time scale, time since the beginning of exposure, can be written as the difference between t and age at the beginning of exposure. Therefore, for presentational simplicity, in this paper, we assume the time scale for assessing susceptibility and latency is the same as that in the Cox regression model.

Our goal is to provide methods for joint estimation and inferences about **a** and β in models (1) and (2). In Section 2, we present commonly used exposure metrics, and in Section 3 we use the Cox model (3) as an example to describe methods for estimation and inferences about the parameters in the exposure metrics. In Section 4 we describe analysis of the NHS air pollution data and a simulation study is given in Section 5. We end with a discussion in Section 6.

2. Exposure metrics

Many time-varying exposure metrics can fit into framework

$$X(\mathbf{c}(t);\mathbf{a})=\int_{t_0}^{t}\omega(s,t;\mathbf{a})c(s)ds, \quad (4)$$

where t_0 is the age at entry into the cohort, and $\omega(s, t; \mathbf{a})$ is a pre-specified real-valued nonnegative weight function. In Table 1, we define some time-varying exposure metrics, which have previously been considered in epidemiologic research. They all fit into framework (4). The continuous versions of the metrics in Table 1 are $X(\mathbf{c}(t); \mathbf{a})$ in models (1), (2) and (3), and the discrete ones, based on the exposure levels observed at a collection of time points, can be used to approximate the continuous ones. Rarely if ever can the continuous metrics be directly used in practice; to do so would require continuous exposure measurements. Instead, discrete versions of these conceptual metrics are what are observable in practice. For each of exposures (i) to (iv), we have a version of moving average exposure, which reflects averaged intensity in the time window, and a version of cumulative exposure,

which reflects total exposure in the time window. For exposure metric (i), the *a*–month (or another time unit) moving average exposure is often used in air pollution epidemiology (e.g., [6, 5]), where *a* defines the beginning of the recent exposure susceptibility period, and the total cumulative exposure, the *a*–month (or another time unit) cumulative exposure is often used for Radon exposure [15]. In (ii), *a* defines the age-related susceptibility. In metrics (iii) and (iv), t_0 is a pre-specified time or age, often age at entry into the cohort. In (v), what matters is only exposure at age *a*. In (vi), what matters is exposure *a* years ago.

In addition to the exposure metrics in Table 1, a flexible model that is useful for exploring time-dependent exposure effects across the entire observed range of the exposure history is

$$\lambda(t)=\lambda_0(t)r\Big(\sum_{S=t_0}^{t} c(s)\beta_\phi(t-s)+\boldsymbol{\beta}_u\mathbf{U}(t)\Big),\ \text{or}\ \lambda(t)=\lambda_0(t)+d\Big(\sum_{S=t_0}^{t} c(s)\beta_\phi(t-s)+\boldsymbol{\beta}_u\mathbf{U}(t)\Big), \quad (5)$$

where $\beta_\varphi(s)$ may depend on a vector of unknown parameters $\boldsymbol{\varphi}$. Using the Cox model $\lambda(t)=\lambda_0(t)\exp\big(\sum_{s=t_0}^{t} c(s)\beta_\phi(t-s)+\boldsymbol{\beta}_u\mathbf{U}(t)\big)$ as an example, $\beta_\varphi(s)$ is interpretable as the logarithm of RR for a one unit increase in exposure received *s* years previously, $c(t-s)$, while fully controlling for the effects of exposure at other time points prior to *t*. The authors of [16] and [17] considered a piecewise constant model where the latency period is reduced to a few intervals, say *k* intervals, with the *i* th interval $(t-t_i,\ t-t_{i+1})$, and the log relative risk is estimated for each interval. This latency model can be written as a special case of the exposure model (5), where $\beta_\varphi(t-s)=\beta_i$ for $t_{i+1}\le s\le t_i$ and $\boldsymbol{\varphi}=(\beta_1,\ldots,\beta_k)$. As noted in [16], this model is intended as a "first pass" model to get a sense of the shape of the latency function. In the bilinear latency model considered in [16] and [18], the exposure effect as a function of time consists of attached straight lines. A common type of the bilinear latency model is characterized by three times, say a_0, a_1 and a_2, on the latency scale. Up to time a_0 there is no effect of exposure. Then, the relative effect increases linearly, reaching a peak a_1 years in the past and decreases linearly thereafter, reaching zero (no effect) at a_2 years in the past. Furthermore, to accommodate the assumption that the effect of exposure would never entirely disappear, the authors in [16] proposed an alternative model that simply replaces the second line by an exponential decay curve.

Other authors modeled the coefficient function $\beta_\varphi(t-s)$ in model (5) using cubic B-splines, allowing the exposure effect to vary arbitrarily with age [19, 20, 21, 22]. Zanobetti and others [23] developed a generalized additive distributed lag model for the estimation of acute air pollution effects, extending a generalized linear model relating the mean outcome to $\sum_{s=t_0}^{t} c(s)\beta_\phi(t-s)$ and other covariates through an appropriate link function, and estimating the curve $\beta_\varphi(t-s)$ through a penalized spline function [24]. Thomas (2009, Chapter 6) [3] discussed additional exposure metrics useful in the study of the effects of radiation, uranium mining, domestic radon exposures and tobacco smoking. Model (5) is more flexible than the pre-specified exposure metrics in Table 1, and through Model (5), we can estimate the exposure effect trajectory over time, in which the value at time *t* on the curve represents the exposure effect at time *t* while adjusting for exposure levels at all other time points. In contrast, methods which make use of pre-specified exposure metrics provide

point and interval estimates of latency parameters and of a regression coefficient that can be interpreted as the overall effect of the exposure in its estimated critical time window of susceptibility. These quantities may be useful for policy making and for making public health recommendations. The bilinear latency model mentioned above, which also estimates parameters of public health relevance, can be seen in between the methods which make use of pre-specified exposure metrics and model (5) that uses splines to model the exposure effect. This paper will focus on the exposure metrics in Table 1, rather than model (5). None of the methods given thus far directly apply to joint estimation of β and **a** for the exposure metrics in Table 1.

3. Estimation and Inference

In this section, we describe the likelihood-based estimation methods for the Cox model (3).

3.1. Point estimation

We generalize the maximum partial likelihood estimator (MPLE) [25] for **a** and $\boldsymbol{\beta} = (\beta, \boldsymbol{\beta}_u)$, where the dimension of **a**, dim(**a**), is 1 for metrics (i, ii, iii.1, iv.1, v, vi) and 2 for metrics (iii.2, iv.2, iv.3). If dim(**a**) = 1, **a** = *a* in Table 1, but in this section, we denote **a** = a_1 for presentational convenience, and if dim(**a**) = 2, **a** = (a_1, a_2). When there are no ties, the partial likelihood is

$$L(\mathbf{a},\boldsymbol{\beta})=\prod_{i\in\Im}\frac{\exp\{\beta X(\mathbf{c}(t_i;i);\mathbf{a})+\boldsymbol{\beta}_u\mathbf{U}_i(t_i)\}}{\sum_j Y_j(t_i)\exp\{\beta X(\mathbf{c}(t_i;j);\mathbf{a})+\boldsymbol{\beta}_u\mathbf{U}_j(t_i)\}},$$

where *i* refers to the *i* th participant, t_i is the event time of the *i* th participant, $\Im$ is a subset containing all the cases, and $Y_i(t)$ is the at-risk process for the *i* th individual, equal to 1 if at risk at time *t* and 0 otherwise. Let $N_i(t)$ be the counting process for the number of observed failures on (0, *t*]. The partial likelihood score function based on data available up to a specified time *t* is

$$D(\mathbf{a},\boldsymbol{\beta})=\sum_{i\in\Im}\int_0^t\left\{\begin{pmatrix}\beta\{\partial X(\mathbf{c}(u;i);\mathbf{a})/\partial\mathbf{a}\}\\ X(\mathbf{c}(u;i);\mathbf{a})\\ \mathbf{U}_i(u)\end{pmatrix}-\frac{S_1(u)}{S_0(u)}\right\}dN_i(u),$$

where

$$S_1(u)=\sum_{i\in\Im}\begin{pmatrix}\beta\{\partial X(\mathbf{c}(u;i);\mathbf{a})/\partial\mathbf{a}\}\\ X(\mathbf{c}(u;i);\mathbf{a})\\ \mathbf{U}_i(u)\end{pmatrix}Y_i(u)\exp\{\beta X(\mathbf{c}(u;i);\mathbf{a})+\boldsymbol{\beta}_u\mathbf{U}_i(u)\},$$

$$S_0(u)=\sum_i Y_i(u)\exp\{\beta X(\mathbf{c}(u;i);\mathbf{a})+\boldsymbol{\beta}_u\mathbf{U}_i(u)\}.$$

When the average exposure metric (i), $X(c(t);a)=\int_{t-a}^{t} c(s)ds/a$, is of interest, it follows that $\partial X/\partial a = c(t-a)/a - X/a$. That is, the partial score function D is a function of a through function $c(\cdot)$. This is also true for the total cumulative exposure metric (i) and metrics (ii–iv). For metrics (v–vi), the partial likelihood is a function of **a** through function **c**(·), and existence of $\partial X(\mathbf{c}(u; t); \mathbf{a})/\partial \mathbf{a}$ requires first-order differentiability of **c**(·). A challenge in estimation and inference for the parameter **a** using the MPLE method is that **c**(t) is observed only at discrete time points and thus $\partial D/\partial \mathbf{a}$ and $E(\partial D/\partial \mathbf{a})$ are unknown or even do not exist. Therefore, the Newton-Raphson approach is not applicable here.

To obtain the point estimates of **a** and **β**, we use a method combining a grid search and the Newton-Raphson approach [16]. Denote the set of all possible values for the latency parameter a_k given the data as A_k, $k = 1$ if dim(**a**) = 1 and $k = 1, 2$ if dim(**a**) = 2. If the exposure history is observed regularly at pre-specified time intervals, which may be, for example, every month, every year, or every two years, the possible range of ak can be from minimum, a_k^l, to the maximum value, a_k^u, permitted by the data; that is, $A_k=[a_k^l, a_k^u]$. For example, if the longest exposure follow-up time in a study is 60 months, and if a_1 is the latency parameter in exposure metric (i), a_1^l and a_1^u can be set to 0 and 60 months, respectively. The proposed method also applies if the exposure is observed at an irregular pattern which may vary by subject. For example, if the exposure is observed monthly for some participants and bimonthly for the other participants, A_k may contain every month from a_k^l to a_k^u. For those participants with only bimonthly exposure measurements, missing indicators in the discrete versions of the average exposure metrics in Table 1 or a carrying-forward approach in Table 1 will take care of the missing data issue. Let $A = A_1$ if dim(**a**) = 1 and $A = \{a_1, a_2; a_1 \in A_1, a_2 \in A_2\}$ if dim(**a**) = 2. For each fixed value or vector of **a** in A, we use the Newton-Raphson method to obtain the MPLE for **β**, denoted as $\hat{\boldsymbol{\beta}}_{\mathbf{a}}$, and denote the profile partial likelihood under $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}_{\mathbf{a}}$ as $PL(\mathbf{a})$; i.e., $PL(\mathbf{a}) = L(\mathbf{a}, \hat{\boldsymbol{\beta}}_{\mathbf{a}})$, where $L(\mathbf{a}, \boldsymbol{\beta})$ is the partial likelihood. The proposed MPLE for **a**, denoted as $\hat{\mathbf{a}}$, is the value of **a** in A that maximizes $PL(\mathbf{a})$; i.e., $\hat{\mathbf{a}} = \text{argmax}\{PL(\mathbf{a}), \mathbf{a} \in A\}$, and the MPLE for **β**, denoted as $\hat{\boldsymbol{\beta}}$, is $\hat{\boldsymbol{\beta}}_{\hat{\mathbf{a}}}$.

3.2. Profile likelihood confidence interval for a

As discussed above, the partial score corresponding to element **a**, for metrics (i–iv), or the partial likelihood, for metrics (v–vi), is a function of **a** through function $c(\cdot)$. Function $c(\cdot)$ is available only at discrete time points of **a** and its closed form as a function of **a** is unknown. Thus, the standard asymptotic variance estimator for $\hat{\mathbf{a}}$ does not apply. We propose to use a profile likelihood confidence interval (CI) method for **a**, where the $1 - \alpha$ CI for **a** is the elements in A satisfying

$$\log PL(\mathbf{a}) \geq \log L(\hat{\mathbf{a}}, \hat{\boldsymbol{\beta}}) - \frac{1}{2}\chi^2_{\dim(\mathbf{a})}(1-\alpha), \quad (6)$$

where χ_q^2 is the cumulative function of χ^2 distribution with q degrees of freedom. If dim(**a**) = 2, the CI above for **a** is the joint CI of a_1 and a_2, and the marginal CI for a_k is

$$\{a_k{:}a_k \in A_k, \log L(a_k, \hat{a}_{ja_k}, \hat{\boldsymbol{\beta}}_{a_k}) \geq \log L(\hat{\mathbf{a}}, \hat{\boldsymbol{\beta}}) - \frac{1}{2}\chi_1^2(1-\alpha)\}, \quad (7)$$

where $\hat{a}_{ja_k}$ is the MPLE estimate of a_j with a_K fixed, $j = 2, 1$ for $k = 1, 2$, respectively. Since the CI for **a** contains only discrete values in the set A, the equalities in formula (6) and (7) are typically not reached, and thus the coverage rates of these $1 - \alpha$ CIs may be smaller than $1 - \alpha$. Define δ_k and η_k such that the CI for a_k can be written as $\{a_k{:}\hat{a}_k - \delta_k \leq a \leq \hat{a}_k + \eta_k\}$. A modified version of the $1 - \alpha$ CI for a_k which has coverage rate at least $1 - \alpha$ is

$$\{a_k{:}a_k \in A_k, \max(\hat{a}_k - \delta_k - 1, a_k^l) \leq a_k \leq \min(\hat{a}_k + \eta_k + 1, a_k^u)\}. \quad (8)$$

We will refer to this CI as the at-least CI. In Appendix A, we show that, if **c**(*t*) is first-order differentiable in metrics (i–iv) and second-order differentiable for metrics (v–vi), the profile partial likelihood CI method is valid for interval estimation of **a**. In the simulation study discussed in Section 5, in which **c**(*t*) was not continuous, the method still performed well. It would require intensive computing to use the profile likelihood method to obtain interval estimates for **β**. In order to minimize the computational burden, we propose an alternative method below.

3.3. Hessian matrix variance for β

A profile CI for **β** as above has the disadvantage of being computationally intensive. Although we might naturally turn, instead, to a Hessian matrix method to estimate the variance of **β** for exposure metrics (i–iv), the expected value of the Hessian matrix $E(\partial D/\partial(\mathbf{a},\boldsymbol{\beta}))$ is unknown because it involves the second order derivative of $X(\mathbf{c}(t); \beta)$ with respect to **a**. Because we proved in Appendix A that $E(D^T D + \partial D/\partial(\mathbf{a},\boldsymbol{\beta})) = 0$ if **c**(*t*) is continuous and first-order differentiable, although observed only at discrete time points, we propose to estimate the variance of $(\hat{\mathbf{a}}, \hat{\boldsymbol{\beta}})$ using $\widehat{\mathrm{var}}\,(\hat{\mathbf{a}}, \hat{\boldsymbol{\beta}}) = \{E(D^T D)\}^{-1}|_{(\mathbf{a},\boldsymbol{\beta})=(\hat{\mathbf{a}},\hat{\boldsymbol{\beta}})}$. Then, Wald-type confidence intervals for **β** can be obtained from these variance estimates.

3.4. When there is no exposure effect

Parameter **a** is undefined if $\beta = 0$. This poses a challenge for testing H_0: $\beta = 0$ since **a** is defined only under the alternative hypothesis. Based on earlier work on the supremum statistic [26, 27] and work by Zheng and Cheng (2005) [28], Zucker, Agami and Spiegelman (2013) [29] considered a type of supremum statistic *SUP*2 for a change point problem in a Cox regression model, which has some similarity to the latency problem considered here when dim(**a**)=1. Specifically, if dim(**a**)=1, $SUP2 = \max(|\Delta(\mathbf{a}^{(1)})|, |\Delta(\mathbf{a}^{(2)})|)$, where $\mathbf{a}^{(1)}$ and $\mathbf{a}^{(2)}$ are the minimum and maximum values for **a**, $\Delta(\mathbf{a}) = D_0(\mathbf{a})/\sqrt{\mathrm{var}(D_0(\mathbf{a}))}$ for $\mathbf{a} = \mathbf{a}^{(1)}$, $\mathbf{a}^{(2)}$, and D_0 is the partial likelihood score statistic for fixed **a** under the null hypothesis, given by

$$D_0(\mathbf{a}) = \sum_i \int_0^t \left\{ X(\mathbf{c}(u;i);\mathbf{a}) - \frac{S_{10}(X(\mathbf{c}(u;i);\mathbf{a}), u)}{S_{00}(u)} \right\} dN_i(u),$$

with $S_{10}(\mathbf{g}, u) = \sum_i g_i Y_i(u) \exp(\boldsymbol{\beta}_u \mathbf{U}_i(u))$, $\mathbf{g}$ being a vector with the ith element g_i, and $S_{00}(u) = \sum_i Y_i(u) \exp(\boldsymbol{\beta}_u \mathbf{U}_i(u))$. Let $S_{20}(\mathbf{g}, \tilde{\mathbf{g}}, u) = \sum_i g_i \tilde{g}_i Y_i(u) \exp(\boldsymbol{\beta}_u \mathbf{U}_i(u))$, where $\tilde{\mathbf{g}}$ is a vector with the i th element $\tilde{g}_i$, and

$$C(\mathbf{g}, \tilde{\mathbf{g}}) = \sum_i \int_0^t \left\{ \frac{S_{20}(\mathbf{g}, \tilde{\mathbf{g}}, u)}{S_{00}(u)} - \frac{S_{10}(\mathbf{g}, u) S_{10}(\tilde{\mathbf{g}}, u)}{S_{00}^2(u)} \right\} dN_i(u).$$

Let $p = \dim(\mathbf{U}_i)$. Define vectors $\mathbf{g}_j$ for $j = 1, \ldots, p$, and $\mathbf{g}_{p+1}(\mathbf{a})$ such that the ith element of $\mathbf{g}_j$ is the jth element of $\mathbf{U}_i(u)$ for $j = 1, \ldots, p$, and the ith element of $\mathbf{g}_{p+1}(\mathbf{a})$ is $X(\mathbf{c}(u; i); \mathbf{a})$. Let $\boldsymbol{\Omega}$ denote the matrix $C(\mathbf{g}_j, \mathbf{g}_k)_{j,k=1,\ldots,p}$, and let $h(\mathbf{a})$ denote the column vector with components $C(\mathbf{g}_j, \mathbf{g}_{p+1}(\mathbf{a}))$, $j = 1, \ldots, p$. We have $\mathrm{var}(D_0(\mathbf{a})) = C(\mathbf{g}_{p+1}(\mathbf{a}), \mathbf{g}_{p+1}(\mathbf{a})) - h(\mathbf{a})^T \Omega^{-1} h(\mathbf{a})$, and the correlation coefficient of $\Delta(\mathbf{a}^{(1)})$ and $\Delta(\mathbf{a}^{(2)})$ can be estimated by $\{C(\mathbf{g}_{p+1}(\mathbf{a}^{(1)}), \mathbf{g}_{p+1}(\mathbf{a}^{(2)})) - h(\mathbf{a}^{(1)})^T \Omega^{-1} h(\mathbf{a}^{(2)})\} \{\mathrm{var}(D_0(\mathbf{a}^{(1)}))\mathrm{var}(D_0(\mathbf{a}^{(2)}))\}^{-1/2}$, where $\boldsymbol{\beta}_u$ is replaced with the MPLE $\hat{\boldsymbol{\beta}}_u$ under the model with $\beta = 0$. The critical values for the *SUP*2 statistics can be obtained using established routines for computing multivariate normal probabilities [28]. This approach can be extended to $\dim(\mathbf{a})=2$ with $\mathbf{a} = (a_1, a_2)$ by including terms in addition to $\Delta(\mathbf{a}^{(1)})$ and $\Delta(\mathbf{a}^{(2)})$ in the test statistic, with $\mathbf{a}^{(1)} = (a_{1min}, a_{2min})$ and $\mathbf{a}^{(2)} = (a_{1max}, a_{2max})$ now, where a_{kmax} and a_{kmin} are the maximum and minimum values of a_k, $k = 1, 2$. For example, for exposure metrics (iii.2) and (iv.3), the test statistic could be the supremum statistic *SUP*3 [28, 29] based on $\max(|\Delta(\mathbf{a}^{(1)})|, |\Delta(\mathbf{a}^{(2)})|, |\Delta(\mathbf{a}^{(3)})|)$, where $\mathbf{a}^{(3)} = (a_{1min}, a_{2max})$.

Now consider the case when data are generated from model (3) with $\beta = 0$. Since now model (3) is true for any given $\mathbf{a}$, we can obtain $\hat{\beta}$ by the average of MPLE $\hat{\beta}_{\mathbf{a}}$ over a set of given values of $\mathbf{a}$; that is, denoting the set of possible values of $\mathbf{a}$ as A,

$$\hat{\beta} = \sum_{\mathbf{a} \in A} \hat{\beta}_{\mathbf{a}} / m, \quad (9)$$

where m is the number of values of $\mathbf{a}$ in A. Using a formula well established in the multiple imputation literature [30], we have

$$\mathrm{var}(\hat{\beta}) = Q + (1 + 1/m) B, \quad (10)$$

where $Q = \sum_{\mathbf{a} \in A} \mathrm{var}(\hat{\beta}_{\mathbf{a}})/m$, and $B = \sum_{\mathbf{a} \in A} (\hat{\beta}_{\mathbf{a}} - \hat{\beta})^2/(m - 1)$.

In practice, since we do not know if β is zero or not, point estimates and confidence intervals for β may be derived based on a supremum statistic-based test. First, test the hypothesis H_0: $\beta = 0$ using the supremum statistic. If the null hypothesis is rejected, the joint MPLE method described in Sections 3.1–3.3 can be used to obtain the point and interval estimate of (β, $\mathbf{a}$); otherwise, the point and interval estimate of β may be obtained using (9) and (10).

4. Illustrative Example

We applied these methods to evaluate the relationship between fine particulate matter less than 2.5 μm in diameter ($PM_{2.5}$) and all-cause mortality extending the analysis in NHS [5],

based on the Cox model (3). NHS began in 1976 with 121,700 female registered nurses aged 30–55 years who completed a mailed questionnaire about their health and lifestyle. At the time of the study's inception, the nurses resided in 11 states throughout the United States. Since that time, participants have moved into all 50 states. In this analysis, the 72 month follow-up period of this study began in July 2000 and ended in June 2006, and following the previous analysis [5], we excluded participants who were living outside metropolitan statistical areas (MSAs) because the distributions of air pollution monitors and nurses were sparse [5]. $PM_{2.5}$ data based on a spatio-temporal model [31] are available monthly from January 1999 to June 2006, while in the previous analysis [5], it is from 1999 to 2002. Table 2 shows the basic characteristics of the study population. During 6,428,433 person-months of follow-up, 6211 deaths from all causes excluding accidental deaths occurred among the 92,140 participants, while in [5] there were 606,752 person-months and 3,785 deaths among 66,250 women.

To illustrate the methods, following [6, 5, 7], we considered the moving average exposure as exemplified by exposure metric (i), where the range of a was [0, 72] and $A = \{0, 1, \ldots, 72\}$. The supremum $SUP2$ test for testing H_0: $\beta = 0$ had a p-value of 0.009, thus a and β were estimated jointly using the maximum partial likelihood method. The point estimate of a based on the MPLE method was 7 months, with the 95% CI of 6 to 8 months. The estimated RR for the effect of the 7-month moving average $PM_{2.5}$ on all-cause mortality was 1.25 per 10 $\mu g/m^3$, with the 95% Hessian-based CI (1.21, 1.30). Here, the strongest effect method produced the same $\hat{a}$ and $\hat{RR}$. In [5], which was based on a shorter follow-up period from 1999 to 2002 and adjusted for more confounders, the 36-month moving average $PM_{2.5}$ corresponded to the strongest effect.

5. Simulation study

We conducted a simulation study to evaluate the finite sample performance of the proposed methods for the recent moving cumulative average exposure (exposure metric i), and the average exposure during a critical period of susceptibility (exposure metric iii.1), based on the Cox model (3). The exposure data were generated following the distribution of monthly $PM_{2.5}$ ($\mu g/m^3$) in the NHS air pollution study, using a multivariate normal distribution with mean 14 $\mu g/m^3$, standard deviation 4 $\mu g/m^3$, and an intra-class correlation coefficient (ICC) for two successive exposure measurements within subject at months $t1$ and $t2$ equal to $0.6^{|t_1-t_2|}$, as found in the data.

The outcome data were generated from the Cox model (3) under the recent moving average exposure (metric i in Table 1) or the average exposure during critical period of susceptibility (metric iii.1 in Table 1) with $\beta = 0.3$ and a =5, 10, 25, and 35 months, and $\beta = 0.0$, following the time to event data generation method described in the Appendix of [32]. The baseline hazard function was assumed to be of Weibull form $\lambda_0(t) = \theta\nu(\nu t)^{\theta-1}$, with $\theta = 6.0$, as is typical of many epithelial cancers [33, 34]. Censoring was assumed exponential with a rate of 1.5% per month. The parameter ν was set to achieve cumulative incidence of 5% and 25%, with 500 replicates for each design point.

Shown in Tables 3 and 4 are the simulation results for the two exposure metrics mentioned above. The $SUP2$ test for testing H_0: $\beta = 0$ had a rejection rate ranging from 0.046 to about 0.06 for a significance level of 0.05 when the data were generated under the null. The supremum $SUP2$ test-based point estimates and confidence intervals of β described in Section 3.4 also performed well. The profile likelihood confidence intervals of a had coverage rates close to 0.95 in most simulation scenarios, except when a was large. For larger a, more events are needed to achieve a good coverage rate. In contrast, the strongest effect a-estimator greatly overestimated a and the strongest effect β-estimates were biased away from the null. We also conducted a simulation with the exposure measurement frequency doubled, and the simulation results for point estimates and confidence intervals, which are not reported in this paper, were similar to those when the exposure measurement frequency was not doubled. In addition, we have conducted a simulation study for the moving average exposuremetric iv.3, with a two-dimensional parameter, $\mathbf{a} = (a_1, a_2)$. This simulation study shows that the method works well for exposure metrics with two dimensional parameters. When $a_1 = 10$, $a_2 = 15$ and $\beta = 0.3$, the mean and median of $\hat{a}_1$ from 500 simulation replicates based on our proposed method were 10.9 and 11 and those of $\hat{a}_2$ were 15.5 and 15, with a sample size of 10, 000; the mean and median of $\hat{a}_1$ improved to 10.3 and 10 and those of $\hat{a}_2$ were 15.1 and 15, with a sample size of 20, 000. The coverage rates of the 95%two-dimensional profile confidence region of $\mathbf{a} = (a_1, a_2)$ calculated based on (6) were 0.97 and 0.95, respectively. In contrast, the mean and median of $\hat{a}_1$ based on the extreme effect method were 5.8 and 6 and those of $\hat{a}_2$ were 22.4 and 22, for a sample size of 10, 000; the mean and median of $\hat{a}_1$ were 5.7 and 6 and those of $\hat{a}_2$ were 21.7 and 21, for a sample size of 20, 000. The means of $\hat{\beta}$ from the test-based method in Section 3.4 were 0.31 for both sample sizes of 10, 000 and 20, 000, and the covariate rates of the 95% CIs of β were 0.86 and 0.91 for sample sizes of 10, 000 and 20, 000, respectively; The means of $\hat{\beta}$ from the extreme effect method were 0.32 for both sample sizes.

6. Discussion

This paper considers time to event data with a time-varying exposure that is a function of the exposure history when a latency period, age-related susceptibility, and other timing of exposure issues characterize the exposure effect. We propose likelihood-based methods for inference on the parameters over a wider range of exposure metrics than previously considered, where the parameters of these metrics may represent the duration of the latency period or age-related susceptibility. Although the methods developed in this paper, the motivating data example, and the simulation study in Sections 3 to 5 are all pertain to the Cox models (3), this likelihood-based estimation framework can be used for both the relative risk model (1) and the excess risk model (2). For the latter, since the baseline incidence rate, λ_0, does not cancel out as in the partial likelihood for a relative risk model, distributional assumptions are needed for λ_0, and a full likelihood approach is required.

A user-friendly publicly available SAS macro and R function are under development, and will be a useful tool for the many studies which collect exposure histories over time, as is common in environmental, occupational, nutritional and life course epidemiology. While the SAS macro and R function are under development, the Fortran program implementing the method is available upon request to the first author.

We considered an alternative method where we first smoothed the exposure trajectory so that standard likelihood-based methods could be applied to jointly estimate the parameters of the latency function and the RR of the exposure, following methods developed in the functional data analysis literature [35, 36]. We used a mixed model method to smooth the exposure trajection [36, 37], and then used standard likelihood methods for inference on **a** and β. Similar methods have been used in Wang and Choi (2014) [38] and Sanchez *et al.* (2011) [39], both of which are for continuous outcomes in the area of prenatal susceptibility to toxicants. In our simulation study, we found that this method converged to the true parameters as the number of events increased at a slower rate than the method proposed in this paper, and thus would typically require an unrealistically large number of events in order to achieve satisfactory finite sample performance, so we do not present this method here. We will investigate this pre-smoothing approach further in our future research.

Although both Langholz *et al.* (1999) [16] and this paper use a grid search over the profile likelihood jointly to find the estimates, Langholz *et al.* (1999) [16] and ours apply to different settings. In [16], the latency parameters are weight functions that do not involve the discrete exposures. In this paper, the latency parameters are in the bounds of the integral of the discretely measured exposure. Our paper advances the methods in the following ways not covered by [16]:We propose profile likelihood confidence intervals for the latency parameters, prove in the Appendix that the proposed method is valid for the partial likelihoods, consider the case when there is no exposure effect and propose a test-based point and interval estimates for **a** and β. In addition, we evaluate all of these methods in an extensive simulation study. Note that the discrete latency models adopted in this paper are biologically not plausible and are approximations to models based on continuous exposure measurements, which are rarely, if ever, observable.

For presentational simplicity, in this paper, we assume the time scale for assessing susceptibility and latency is the same as that used in the Cox regression model. When the two time scales are different, the time scale for assessing susceptibility and latency can be typically transformed to that in the Cox model. How to choose the exposure metric may be based on biological knowledge, empirical methods [40, 41] and convention, and is beyond the scope of this paper. We refer the reader to a useful discussion in Section “Extended exposure histories” in Chapter 6 of [3] about selection between a moving average exposure metric and a total cumulative exposure metric in relation to the choice between the relative risk model versus the excess risk model when the disease outcomes are cancers. Topics for future research include methods for obtaining valid point and interval estimates for **a** and β in the presence of exposure measurement error, and power considerations for the joint estimation of latency parameters and exposure effects, as a function of the sample size, event rate, exposure ICC, and other features for common exposure metrics. It appears that much larger studies are needed to estimate these more complex models with adequate power.

Acknowledgments

We thank the Associate Editor and two referees for their helpful comments that improved the paper. This project was supported by Grant R01 ES009411-09 from the National Institute of Environmental Health Science (NIEHS), National Institutes of Health.

APPENDIX A

Standard asymptotic theory for the Cox model was proven when the unknown parameters are all regression coefficients [42, 43]. Here, we show that when a parameter is not a regression coefficient (e.g., the latency parameter *a* in the average exposure metric i), the profile likelihood confidence interval method is still valid. The proof is similar to that given in [42] and [43] for regression coefficients.

Model (3) can be written in the following general form $\lambda_i(t) = \lambda_0(t)\exp(F_i(t;\zeta))$, where *i* refers to the *i* th individual, and ζ is a vector of unknown parameters including those in the exposure metrics, **a**, β and β_u. Assume $c(t)$ is first-order differentiable if using metrics (i–iv) and second-order differentiable for metrics (v–vi). It follows that $F(t;\zeta)$ is second-order differentiable with respect to ζ. Let F' and F'' denote the first and second derivatives of F with respect to ζ, respectively. In the case of no ties, the partial likelihood score function based on data available up to a specified time *t* is

$$D(\zeta)=\sum_i \int_0^t \{F_i'(u) - \frac{S_1(u)}{S_0(u)}\} dN_i(u),$$

where $S_1(u)=\sum_i F_i'(u)Y_i(u)\exp(F_i(u))$ and $S_0(u)=\sum_i Y_i(u)\exp(F_i(u))$.

Below we will show that (i) the partial score function has zero mean; it will then follow that the MPLE of ζ is consistent; (ii) $E(D^T D + D') = 0$. Under these conditions, it is straightforward to show that the likelihood ratio test is valid and thus the profile likelihood confidence interval method applies.

Proof of (i):

The compensator of $N_i(t)$ is $A_i(t)=\int_0^t Y_i(u)\exp(F_i(u))\lambda_0(u)du$. By simple algebra, we have

$$\sum_i \int_0^t \{F_i'(u) - \frac{S_1(u)}{S_0(u)}\} dA_i(u)=0.$$

It follows that

$$D(\zeta)=\sum_i \int_0^t \{F_i'(u) - \frac{S_1(u)}{S_0(u)}\} M_i(u),$$

where $M_i(u)$ is a zero mean martingale. The *i*th term is a stochastic integral of a predictable vector process with respect to a martingale. Thus, *D* is itself a mean 0 vector-valued martingale. This proves the consistency of the ζ -estimator.

Proof of (ii):

First $D'(\zeta)$ can be written as sum of two terms, D_1 and D_2, where

$$D_1=\sum_i \int_0^t \{F''(u) - \frac{S_2(u)}{S_0(u)}\} dN_i(u),$$

$$D_2=\sum_i \int_0^t \{-\frac{S_3(u)}{S_0(u)}+\frac{S_1^T(u)S_1(u)}{S_0^2(u)}\} dN_i(u),$$

where $S_2(u)=\sum_i F_i''(u)Y_i(u)\exp(F_i(u))$ and $S_3(u)=\sum_i F_i'(u)^T F'(u)Y_i(u)\exp(F_i(u))$. Similar to the argument used for proving the unbiasedness of $D(\zeta)$, D_1 has zero mean. Note that

$$D^T D=\sum_i \int_0^t \{F_i'(u)^T F_i'(u) - \frac{2F_i'(u)^T S_1(u)}{S_0(u)}+\frac{S_1^T(u)S_1(u)}{S_0^2(u)}\} dN_i(u).$$

Consider $D^T D + D'$. Since $\sum_i \int_0^t \{F_i'(u)^T F_i'(u) - \frac{S_3(u)}{S_0(u)}\} dN_i(u)$ is mean zero due to arguments similar to those used for proving the unbiasedness of $D(\zeta)$, we have

$$E(D^T D+D')=E\left[2\sum_i \int_0^t \{-\frac{F_i'(u)^T S_1(u)S_0(u)+S_1^T(u)S_1(u)}{S_0^2(u)}\} dN_i(u)\right].$$

Since $\sum_i \int_0^t \{-\frac{F_i'(u)^T S_1(u)S_0(u)+S_1^T(u)S_1(u)}{S_0^2(u)}\} dA_i(u)=0$, it follows that

$$E(D^T D+D')=E\left[2\sum_i \int_0^t \{-\frac{F_i'(u)^T S_1(u)S_0(u)+S_1^T(u)S_1(u)}{S_0^2(u)}\} dM_i(u)\right].$$

The i th term is a stochastic integral of a predictable vector process with respect to a martingale. Thus, D is itself a mean 0 vector valued martingale. This proves (ii).

References

1. Preston DL, Kusumi S, Tomonaga M, Izumi S, Ron E, Kuramoto A, Kamada N, Dohy H, Matsuo T, Matsui T. corrected to Matsuo T. Cancer incidence in atomic bomb survivors. part iii. leukemia, lymphoma and multiple myeloma, 1950–1987. Radiat Res. 1994 Feb; 137(2 Suppl):S68–S97. [PubMed: 8127953]
2. Preston DL, Pierce DA, Shimizu Y, Cullings HM, Fujita S, Funamoto S, Kodama K. Effect of recent changes in atomic bomb survivor dosimetry on cancer mortality risk estimates. Radiat Res. 2004 Oct; 162(4):377–389. [PubMed: 15447045]
3. Thomas, D. Statistical Methods in Environmental Epidemiology. New York: Oxford University Press Inc.; 2009.
4. De Stavola B, Nitsch D, Silva I, McCormack V, Hardy R, Mann V, Cole T, Morton S, Leon D. Statistical issues in life course epidemiology. American Journal of Epidemiology. 2006; 163:84–96. [PubMed: 16306313]

5. Puett R, Hart J, Yanosky J, Paciorek C, Schwartz J, Suh H, Speizer F, Laden F. Chronic fine and coarse particulate exposure, mortality, and coronary heart disease in the Nurses' Health Study. Environmental Health Perspectives. 2009; 117:1697–1701. [PubMed: 20049120]
6. Laden F, Schwartz J, Speizer F, Dochery D. Reduction in fine particulate air pollution and mortality: extended follow-up of the harvard six cities study. American Journal of Respiratory and Critical Care Medicine. 2006; 173:667–672. [PubMed: 16424447]
7. Schwartz J, Coull B, Laden F, Ryan L. The effect of dose and timing of dose on the association between airborne particles and survival. Environmental Health Perspectives. 2008; 116:64–69. [PubMed: 18197301]
8. Magnusson C, Baron J, Persson I, Wolk A, Bergstrom R, Trichopoulos D, Adami HO. Body size in different periods of life and breast cancer risk in post-menopausal women. Int. J. Cancer. 1998; 76:29–34. [PubMed: 9533758]
9. Rothman K. Induction and latent periods. American Journal of Epidemiology. 1981; 114:253–259. [PubMed: 7304560]
10. Salvan A, Stayner L, Steenland K, Smith R. Selecting an exposure lag period. Epidemiology. 1995; 6:387–396. [PubMed: 7548346]
11. Richardson D, Cole S, Chu H, Langholz B. Lagging exposure information in cumulative exposure-response analyses. American Journal of Epidemiology. 2011; 174:1416–1422. [PubMed: 22047823]
12. Breslow N, Lubin J, Marek P, Langholz B. Multiplicative models and cohort analysis. Journal of the American Statistical Association. 1983; 78:1–12.
13. Korn E, Graubard B, Midthune D. Time-to-event analysis of longitudinal follow-up of a survey: choice of the time-scale. American Journal of Epidemiology. 1997; 145:72–80. [PubMed: 8982025]
14. Commenges D, Letenneur L, Joly P, Alioum A, Dartigues J. Modelling age-specific risk: application to dementia. Statistics in Medicine. 1998; 17:1973–1988. [PubMed: 9777690]
15. Field R, Steck D, Smith B, Brus C, Fisher E, Neuberger J, Platz C, Robinson R, Woolson R, Lynch C. Residential radon gas exposure and lung cancer. American Journal of Epidemiology. 2000; 151:1091–1102. [PubMed: 10873134]
16. Langholz B, Thomas D, Xiang A, Stram D. Latency analysis in epidemiologic studies of occupational exposures: application to the Colorado Plateau uranium miners cohort. American Journal of Industrial Medicine. 1999; 35:246–256. [PubMed: 9987557]
17. Finkelstein M. Use of "time windows" to investigate lung cancer latency intervals at an Ontario steel plant. American Journal of Internal Medicine. 1991; 19:229–235.
18. Richardson D, MacLehose R, Langholz B, Cole S. Hierarchical latency models for dose-time-response associations. American Journal of Epidemiology. 2011; 173:695–702. [PubMed: 21303803]
19. Hauptmann M, Pohlabeln H, Lubin JH, Jockel K, Ahrens W, Bruske-Hohlfeld I, Wichmann HE. The exposure-time-response relationship between occupational asbestos exposure and lung cancer in two German case-control studies. American Journal of Industrial Medicine. 2002; 41:89–97. [PubMed: 11813213]
20. Hauptmann M, Berhane K, Langholz B, Lubin J. Using splines to analyse latency in the Colorado Plateau uranium miners cohort. Journal of Epidemiology and Biostatistics. 2001; 6:417–424. [PubMed: 11831677]
21. Hauptmann M, Wellmann J, Lubin J, Rosenberg P, Kreienbrock L. Analysis of exposure-time-response relationships using a spline weight function. Biometrics. 2000; 56:1105–1108. [PubMed: 11129467]
22. Sylvestre M, Abrahamowicz M. Flexible modeling of the cumulative effects of time-dependent exposures on the hazard. Statistics In Medicine. 2009; 28(27):3437–3453. [PubMed: 19708037]
23. Zanobetti A, Wand MP, Schwartz J, Ryan LM. Generalized additive distributed lag models: quantifying mortality replacement. Biostatistics. 2000; 1:279–292. [PubMed: 12933509]
24. Eilers P, Marx B. Flexible smoothing with B-splines and penalties (with discussion). Statistical Sciences. 1996; 89:89–121.

25. Cox D. Regression models and life tables(with discussion). Journal of the Royal Statistical Society B. 1972; 34:187–220.
26. Davies R. Hypothesis testing when a nuisance parameer is present only under the alternative. Biometrika. 1977; 64:247–254.
27. Davies R. Hypothesis testing when a nuisance parameer is present only under the alternative. Biometrika. 1987; 74:33–34.
28. Zheng G, Chen Z. Comparison of maximum statistics for hypothesis testing when a nuisance parameter is present only under the alternative. Biometrics. 2005; 61:254–258. [PubMed: 15737101]
29. Zucker DM, Agami S, Spiegelman D. Testing for a changepoint in the Cox survival regression model. Journal of Statistical Theory and Practice. 2013; 7:360–380.
30. Rubin, D. Multiple Imputation for Nonresponse in Surveys. New York: J. Wiley & Sons; 1987.
31. Yanosky J, Paciorek C, Suh H. Predicting chronic fine and coarse particulate exposures using spatio-temporal models for the northeastern and midwestern us. Environmental Health Perspectives. 2009; 117:522–529. [PubMed: 19440489]
32. Liao X, Zucker D, Li Y, Spiegelman D. Survival analysis with error-prone time varying covariates: a risk set calibration approach. Biometrics. 2011; 67:50–58. [PubMed: 20486928]
33. Armitage, P.; Doll, R. Stochastic models for carcinogenesis. In: Newman, J., editor. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability. Berkeley: University of California Press; 1961.
34. Breslow, N.; Day, NE. Statistical Methods in Cancer Reseasrch. World Health Organization; 1993.
35. Muller H, Yao F. Functional additive models. Journal of the American Statistical Association. 2008; 103:1534–1544.
36. Wand MP. Smoothing and mixed models. Computational Statistics. 2003; 18:223–250.
37. Ngo L, Wand MP. Smoothing and mixed model software. Journal of Statistical Software. 2004; 9
38. Wang L, Choi H. Using semiparametric-mixed model and functional linear model to detect vulnerable prenatal window to carcinogenic polycyclic aromatic hydrocarbons on fetal growth. Biometrical Journal. 2014; 56:243–255. [PubMed: 24243524]
39. Sanchez BN, Hu H, Litman HJ, Tellez-Rojo MM. Statistical methods to study timing of vulnerability with sparsely sampled data on environmental toxicants. Environmental Health Perspectives. 2011; 119:409–415. [PubMed: 21362588]
40. Thomas D. General relative-risk models for survival time and matched case-control analysis. Biometrics. 1981; 37:673–686.
41. Hoeting J, Madigan D, Raftery A, Volinsky C. Bayesian model averaging: a tutorial. Statistical Science. 1999; 14:382–417.
42. Kalbfleisch, J.; Prentice, R. The Statistical Analysis of Failure Time Data. Second. Wiley; 2002.
43. Fleming, T.; Harrington, D. Counting Processes and Survival Analysis. Wiley; 1991.

Table 1

Exposure metrics

	$X(c(t);a)$			
	Continuous		**Discrete**	
Metric	**Average**	**Total**	**Average**	**Total**
(i)Moving recent exposure	$\frac{\int_{t-a}^{t} c(s)ds}{a}$	$\int_{t-a}^{t} c(s)ds$	$\frac{\sum_{s=t-a}^{t} I(s)c(s)}{\sum_{s=t-a}^{t} I(s)}$ or $\frac{\sum_{s=t-a}^{t} \tilde{c}(s)}{a+1}$	$\sum_{s=t-a}^{t} \tilde{c}(s)$
(ii)mid- or later-life-related susceptibility window	$\frac{\int_{t-a}^{t} c(s)ds}{t-a}$	$\int_{a}^{t} c(s)ds$	$\frac{\sum_{s=a}^{t} I(s)c(s)}{\sum_{s=a}^{t} I(s)}$ or $\frac{\sum_{s=a}^{t} \tilde{c}(s)}{t-a+1}$	$\sum_{s=a}^{t} \tilde{c}(s)$
(iii)Exposure during critical	1. $\frac{\int_{t_0}^{a} c(s)ds}{a-t_0}$	$\int_{t_0}^{a} c(s)ds$	$\frac{\sum_{s=t_0}^{a} I(s)c(s)}{\sum_{s=t_0}^{a} I(s)}$ or $\frac{\sum_{s=t_0}^{a} \tilde{c}(s)}{a-t_0+1}$	$\sum_{s=t_0}^{a} \tilde{c}(s)$
period of susceptibility	2. $\frac{\int_{a_1}^{a_2} c(s)ds}{a_2-a_1}$	$\int_{a_1}^{a_2} c(s)ds$	$\frac{\sum_{s=a_1}^{a_2} I(s)c(s)}{\sum_{s=a_1}^{a_2} I(s)}$ or $\frac{\sum_{s=a_1}^{a_2} \tilde{c}(s)}{a_2-a_1+1}$	$\sum_{s=a_1}^{a_2} \tilde{c}(s)$
(iv)Age- or time related moving	1. $\frac{\int_{t_0}^{t-a} c(s)ds}{t-a-t_0}$	$\int_{t_0}^{t-a} c(s)ds$	$\frac{\sum_{s=t_0}^{t-a} I(s)c(s)}{\sum_{s=t_0}^{t-a} I(s)}$ or $\frac{\sum_{s=t_0}^{t-a} \tilde{c}(s)}{t-a-t_0+1}$	$\sum_{s=t_0}^{t-a} \tilde{c}(s)$
exposure with a lag	2. $\frac{\int_{a_1}^{t-a_2} c(s)ds}{t-a_2-a_1}$	$\int_{a_1}^{t-a_2} c(s)ds$	$\frac{\sum_{s=a_1}^{t-a_2} I(s)c(s)}{\sum_{s=a_1}^{t-a_2} I(s)}$ or $\frac{\sum_{s=a_1}^{t-a_2} \tilde{c}(s)}{t-a_2-a_1+1}$	$\sum_{s=a_1}^{t-a_2} \tilde{c}(s)$
	3. $\frac{\int_{t-a_1}^{t-a_2} c(s)ds}{a_1-a_2}$	$\int_{t-a_1}^{t-a_2} c(s)ds$	$\frac{\sum_{s=t-a_1}^{t-a_2} I(s)c(s)}{\sum_{s=t-a_1}^{t-a_2} I(s)}$ or $\frac{\sum_{s=t-a_1}^{t-a_2} \tilde{c}(s)}{a_1-a_2+1}$	$\sum_{s=t-a_1}^{t-a_2} \tilde{c}(s)$
(v)One time exposure effect			$c(a)$	
(vi)A lag type model			$c(t-a)$	

$I(s)$ is a missing indicator; it is 1 if $c(s)$ is available and 0 otherwise. $c(\tilde{s})$ is $c(s)$ when $c(s)$ is available, and is the exposure value $c(s$ (carrying backward imputation) or $c(s-\Delta_b)$ (carrying forward imputation) if $c(s)$ is not available and the closest time points when exposure is available is $s-\Delta_f$ and $s+\Delta_b$ in the forward and backward directions, respectively. Assumes missingness is random. There are the following restrictions: for (i), $a \geq 0$; for (ii), $0 \leq a \leq t$; for (iii.1), $a \geq t_0$; for (iii.2), $a_2 \geq a_1 \geq 0$; for (iv.1), $t \geq a + t_0$; for (iv.2), $t \geq a_1 + a_2$; for (iv.3), $a_2 \leq a_1$; for (v) and (vi), $a \geq 0$.

Table 2

The NHS air pollution study (n = 92,140)

N of cases (person-months)	6,211 (6,428,433)
Follow up period (month/year)	June, 2000–June, 2006
Age at study entry (years):	
Median (range)	65 (53,86)
Region:	
Northeast	50%
Midwest	17%
West	14%
South	19%
Monthly $PM_{2.5}$($\mu g/m^3$):	
Mean (sd)	13.7 (4.0)
ICC	0.40

Table 3

Simulation results for moving average exposure (exposure metric i) (500 replicates)

					$\hat{\beta}$				$\hat{a}$			
					MLE†		Str-Eff*		MLE			Str-Eff*
β	*a*	Sample Size	Event %#	Test %reject	mean	CR	mean	mean	median(25th,75th)**	CR	mean	median(25th,75th)**
0.3	5	10,000	25%	0.488	0.302	0.94	0.336	7.0	6 (6, 7)	0.93	24.3	20 (8, 41)
		50,000		0.992	0.307	0.90	0.317	6.3	6 (6, 7)	0.93	19.9	11 (8, 32)
		100,000		1.000	0.306	0.86	0.312	6.2	6 (6, 6)	0.94	15.0	9 (8, 11)
		50,000	5%	0.486	0.300	0.95	0.332	7.0	6 (6, 7)	0.92	22.4	20 (9, 41)
		100,000		0.768	0.304	0.95	0.324	6.6	6 (6, 7)	0.95	22.9	14 (8, 40)
	10	10,000	25%	0.356	0.300	0.97	0.335	12.7	11 (10, 13)	0.90	27.5	24 (13, 42)
		50,000		0.960	0.305	0.94	0.318	11.3	11 (11, 12)	0.96	26.9	22 (14, 41)
		100,000		0.998	0.305	0.93	0.312	11.2	11 (11, 11)	0.96	23.9	16 (13, 36)
		50,000	5%	0.344	0.299	0.95	0.333	12.7	11 (10, 13)	0.92	28.4	27 (14, 42)
		100,000		0.628	0.300	0.95	0.324	12.0	11 (11, 13)	0.92	27.6	23 (14, 43)
	25	10,000	25%	0.322	0.294	0.92	0.324	25.0	25 (18, 32)	0.82	37.0	38 (28, 47)
		50,000		0.948	0.304	0.92	0.312	26.8	26 (24, 28)	0.87	37.5	37 (29, 46)
		100,000		1.000	0.306	0.95	0.312	26.1	26 (25, 27)	0.93	37.2	35 (29, 47)
		50,000	5%	0.350	0.299	0.91	0.327	25.6	26 (20, 32)	0.82	36.6	37 (28, 48)
		100,000		0.612	0.303	0.92	0.320	27.1	26 (23, 31)	0.87	37.7	37 (29, 48)
	35	10,000	25%	0.308	0.292	0.89	0.320	28.4	32 (18, 38)	0.87	39.8	41 (35, 49)
		50,000		0.958	0.305	0.90	0.311	35.8	36 (33, 38)	0.88	42.3	42 (37, 48)
		100,000		1.000	0.302	0.96	0.307	35.8	36 (34, 37)	0.92	43.0	42 (38, 49)
		50,000	5%	0.350	0.299	0.91	0.327	25.6	26 (20, 32)	0.82	36.6	37 (28, 48)
		100,000		0.634	0.299	0.89	0.314	33.0	35 (28, 40)	0.84	42.0	43 (37, 48)
0.0		10,000	25%	0.052	0.004	0.95	0.010			N/A		
		50,000		0.046	−0.001	0.96	0.000			N/A		
		100,000		0.047	−0.002	0.97	−0.002			N/A		
		50,000	5%	0.060	0.002	0.94	0.007			N/A		
		100,000		0.060	0.002	0.94	0.005			N/A		

Time range is (0, 50); exposure is available at $t = 0, 1, \ldots, 50$, as well as at the same number of time points in 50 months before the baseline; these historical exposure data are needed to calculate the moving average exposure; CRs are the empirical coverages of the 95% CIs; CR for a is based on the at-least 95% CI given by expression (8). The 95% CI of the empirical 95% coverage rate in 500 replicates is (.93,.97).

† Means are based on the test-based method in Section 3.4.

Event cumulative rate.

* Strongest effect method.

** Median, 25th percentile, and 75th percentile of $\hat{a}$ from 500 simulation replicates.

Stat Med. Author manuscript; available in PMC 2017 June 15.

Table 4

Simulation results for average exposure during critical period of susceptibility in the past (exposure metric iii.1) (500 replicates)

					$\hat{\beta}$				$\hat{a}$			
					MLE[†]		Str-Eff[*]		MLE			Str-Eff[*]
β	*a*	Sample Size	Event %[#]	Test %reject	mean	CR	mean	mean	median(25th,75th)[**]	CR	mean	median(25th,75th)[**]
0.3	5	10,000	25%	0.760	0.301	0.95	0.326	5.7	5 (4, 6)	0.95	20.5	13 (7, 34)
		50,000		1.000	0.300	0.94	0.308	5.0	5 (5, 5)	0.97	15.6	8 (6, 22)
		100,000		1.000	0.300	0.95	0.305	5.0	5 (5, 5)	0.96	12.5	7 (6, 11)
		50,000	5%	0.802	0.302	0.95	0.326	5.6	5 (4, 6)	0.94	21.7	17 (7, 35)
		100,000		0.984	0.302	0.94	0.317	5.2	5 (5, 5)	0.96	19.4	11 (7, 32)
	10	10,000	25%	0.352	0.300	0.97	0.335	12.7	11 (10, 13)	0.90	27.5	24 (13, 42)
		50,000		0.960	0.305	0.94	0.318	11.3	11 (11, 12)	0.96	26.9	22 (14, 41)
		100,000		0.998	0.305	0.92	0.312	11.2	11 (11, 11)	0.96	23.9	16 (13, 36)
		50,000	5%	0.342	0.299	0.95	0.333	12.7	11 (10, 13)	0.92	28.4	27 (14, 42)
		100,000		0.632	0.300	0.95	0.324	12.0	11 (11, 13)	0.92	27.6	23 (14, 43)
	25	10,000	25%	0.400	0.299	0.93	0.321	24.6	25 (19, 32)	0.85	33.8	33 (26, 42)
		50,000		0.984	0.303	0.93	0.308	25.6	25 (23, 27)	0.91	34.5	32 (27, 42)
		100,000		1.000	0.302	0.95	0.307	25.1	25 (24, 26)	0.96	34.2	32 (27, 40)
		50,000	5%	0.422	0.302	0.92	0.325	25.4	25 (19, 31)	0.88	35.2	35 (27, 43)
		100,000		0.746	0.302	0.92	0.314	25.6	25 (22, 29)	0.88	34.7	33 (27, 42)
	35	10,000	25%	0.310	0.292	0.89	0.320	28.4	32 (18, 38)	0.87	39.8	41 (35, 49)
		50,000		0.954	0.305	0.90	0.311	35.8	36 (33, 38)	0.88	42.3	42 (37, 48)
		100,000		1.00	0.304	0.93	0.308	35.9	36 (34, 38)	0.89	43.2	43 (38, 49)
		50,000	5%	0.344	0.299	0.91	0.327	25.6	26 (20, 32)	0.82	36.6	37 (28, 48)
		100,000		0.638	0.299	0.89	0.314	33.0	35 (28, 40)	0.84	42.0	43 (37, 48)
0.0		10,000	25%	0.046	0.002	0.97	0.005			N/A		
		50,000		0.058	0.000	0.95	0.000			N/A		
		100,000		0.055	0.000	0.95	0.000			N/A		
		50,000	5%	0.046	−0.001	0.97	0.001			N/A		
		100,000		0.053	−0.002	0.97	−0.002			N/A		

Time range is (0, 50), and exposure is available at $t = 0, 1, \ldots, 50$. For notations in the table header, see the footnotes of Table 3.