Harvard University
Harvard University Biostatistics Working Paper Series

Year 2009 *Paper* 110

Survival Analysis with Error-prone Time-varying Covariates: A Risk Set Calibration Approach

Xiaomei Liao* David M. Zucker†
Yi Li‡ donna spiegelman**

*Harvard School of Public Health, stxia@channing.harvard.edu
†Hebrew University
‡Harvard University and Dana Farber Cancer Institute, yili@jimmy.harvard.edu
**stdls@hsph.harvard.edu

http://biostats.bepress.com/harvardbiostat/paper110

Survival analysis with error-prone time-varying covariates: a risk set calibration approach

Xiaomei Liao[1,2,*], David M. Zucker[3], Yi Li[1], and Donna Spiegelman[1,2,]**

[1]Department of Biostatistics, Harvard School of Public Health, Boston, MA, 02115,U.S.A.

[2]Department of Epidemiology, Harvard School of Public Health, Boston, MA, 02115,U.S.A.

[3]Department of Statistics, Hebrew University, Mt. Scopus, 91905 Jerusalem, Israel.

*_email:_ stxia@channing.harvard.edu.

**_email:_ stdls@channing.harvard.edu.

Summary:

Occupational, environmental, and nutritional epidemiologists are often interested in estimating the prospective effect of time-varying exposure variables such as cumulative exposure or cumulative updated average exposure, in relation to chronic disease endpoints such as cancer incidence and mortality. From exposure validation studies, it is apparent that many of the variables of interest are measured with moderate to substantial error. Although the ordinary regression calibration approach is approximately valid and efficient for measurement error correction of relative risk estimates from the Cox model with time-independent point exposures when the disease is rare, it is not adaptable for use with time-varying exposures. By re-calibrating the measurement error model within each risk set, a risk set regression calibration method is proposed for this setting. An algorithm for a bias-corrected point estimate of the relative risk using an RRC approach is presented, followed by the derivation of an estimate of its variance, resulting in a sandwich estimator. Emphasis is on methods applicable to the main study/external validation study design, which arises in important applications. Simulation studies under several assumptions about the error model were carried out, which demonstrated the validity and efficiency of the method in finite samples. The method was applied to a study of diet and cancer from Harvard's Health Professionals Follow-up Study (HPFS).

Key words: Cox proportional hazards model, Measurement error, Risk set regression calibration, Time-varying covariates.

1. Introduction

Many epidemiological studies involve survival data with covariates measured with error: the true covariate value c, as defined by some "gold standard", is represented approximately by a surrogate measure C. Often, interest centers on cumulatively updated total or cumulatively updated average levels of a time-varying exposure, which are computed from a series of error-prone point exposure measurements. For example, in prospective studies of diet and health such as the Nurses' Health Study (Hunter et al., 1996), the primary exposure variable is the cumulatively updated average dietary intake of a given nutrient. Similarly, in prospective studies of the effects of air pollution on health, there is often interest in the effect of cumulative exposure to specific pollutants (Zanobetti et al., 2000). Typically the surrogate point exposures are measured once at each point of a specified grid, and are validated at timepoints in a coarser grid (e.g., Willett et al., 1985; Brauer et al., 2003). There is a practical need for statistical methods suited specifically for such applications.

Covariate measurement error in the Cox survival regression model was first addressed by Prentice (1982), in the setting of a time-invariant exposure. Under certain assumptions, with a linear Gaussian model for c given C, the regression calibration estimator emerged. In the Cox model, the regression calibration estimator is not consistent, but it is a good approximation under certain conditions. In later papers by many authors, more accurate methods were developed for various settings; see Zucker (2005) for a review. We note in particular the risk set regression calibration estimator, which Xie et al. (2001) developed in the setting of a time-invariant exposure under a main/reliability study design. Xie et al. (2001) assumed the classical homoscedastic measurement error model $C = c+\epsilon$, with $E(\epsilon) = 0$ and $Var(\epsilon)$ constant.

Time-varying covariates are more challenging to handle. A number of papers have dealt with measurement error in time-varying covariates. Huang and Wang (2000) presented an

elegant solution for the setting where the classical homoscedastic error model applies and replicate measurements of the surrogate covariate are available on all (or a sizeable sample of) study individuals at *all* times at which events occur.

In practice, as noted above, measurements on the surrogate are usually available only on an intermittent basis. A common strategy is to use the last available covariate measurements, although this strategy can lead to some bias (Raboud et al., 1993). A number of authors have developed more sophisticated methods, based on the joint modeling paradigm. A joint model consists of a model for the covariate process (often a mixed linear model) and a model for the hazard of the relevant event given the covariate (typically a Cox model). Considerable work along this line has been published; Tsiatis and Davidian (2004) provide a recent review.

The joint modeling approach has a number of features that impede its use in applications. The approach is computationally intensive. In addition, it puts an undesirable focus on modeling the exposure process, which requires significant effort but is of no intrinsic interest to the investigators. Moreover, model checking is problematic, because the covariate measurement times are typically too few and too sparse for effective model checking.

As we stated at the outset, we are specifically interested in epidemiological applications involving cumulatively updated total or average exposure. The Willett's classic textbook on nutritional epidemiology cities hundreds of papers which use the cumulatively updated average exposure variable in survival data models, and, similarly, the environmental epidemiology textbooks by Thomas and by Savitz and Steenland cite hundreds of papers using the cumulative exposure and distributed lagged exposure variable in survival data models (Willett, 1998; Thomas, 2009; Steenland and Savitz, 1997). Commonly, in these studies, the point exposures are subject to considerable measurement error, while the error induced by carrying forward the most recent cumulative exposure value is less serious. This motivates an effort to develop methods for analyzing the effect of cumulative exposure in the presence

of measurement error in the point exposures, without invoking the complex joint modeling approach. Methodology of this sort would have immediate applicability in a wide range of large-scale epidemiological studies, including in our own work Harvard's Nurses' Health Studies, the Harvard Professionals Follow-Up Study, the Harvard Six Cities Study, and many others. It would allow cumulative exposure effects to be assessed in a practical way that meets the needs of the applied reality.

The purpose of this paper is to develop such a method. Our approach is to extend the risk set regression calibration method of Xie et al. (2001) from the setting of a time-invariant covariate with a classical measurement model to the setting of cumulative exposure with respect to a time-varying covariate. We work with a measurement error model that is substantially more general than the classical model, and our method is appropriately designed to handle this more complex error structure. Instead of the replicate measures setting, we work under the main study/validation study design, which is suitable for studies in nutritional and environmental epidemiology, where a gold standard measure of exposure, or an unbiased measure thereof, often exists.

Section 2 presents notation and background. Section 3 presents the method. In Section 4, we derive the variance of the proposed estimator for the case of the main study / external validation design. Section 5 presents simulation studies of the method for a range of scenarios motivated by practical applications, including time-varying exposures with different correlation structures, and rare and common disease settings. In Section 6, we illustrate the method on data from the Health Professionals' Follow-Up Study (HPFS) concerning the relationship between total calcium intake and risk of fatal prostate cancer. Section 7 provides a discussion.

2. Definition and preliminary results

The Cox model (Cox, 1972) for censored survival data specifies the hazard rate $\lambda(t)$ for the survival time T of an individual with s-dimensional covariate vector $\mathbf{c}$ to have the form

$$\lambda(t;\mathbf{c}) = \lambda_0(t)\exp\{\boldsymbol{\beta}^t\mathbf{c}\}, \quad t \geqslant 0, \tag{1}$$

where $\boldsymbol{\beta}$ is a s-vector of regression coefficients and $\lambda_0(t)$ is the underlying hazard function.

In the survival data setting with time-invariant covariates, a main/external validation study design consists of data $\{\mathbf{C}_i, \mathbf{W}_i, T_i, D_i\}$, $i = 1, \cdots, n_1$ in the main study, and $\{\mathbf{c}_i, \mathbf{C}_i, \mathbf{W}_i, T_i\}$, $i = n_1 + 1, \cdots, n$ in the validation study. Because data on the outcome, D_i is not available in the validation study, we call this an *external* validation study. Here, $\mathbf{c}$ is the p_1-vector of true exposure which is subject to measurement error, and, in the main study, we observe a vector of surrogate variables $\mathbf{C}$ instead. $\mathbf{W}$ is a p_2-vector of error-free covariates . T is the follow-up time, which is defined as the minimum of the potential failure time T^0 and potential censoring time V, i.e. $T = \min(T^0, V, t^*)$, where t^* is the end of follow up; D is an indicator for failure from the event of interest, n_1 is the sample size of the main study, n_2 is the sample size of the validation study, and $n = n_1 + n_2$. Typically, $\mathbf{c}$ is expensive to measure, and hence $n_1 >> n_2$. In what follows, we start by reviewing the ordinary regression calibration method for several different error models.

Prentice (1982) shows that if $\lambda(t|\mathbf{c}, \mathbf{W}) = \lambda_0(t)\exp(\boldsymbol{\beta}_1^t\mathbf{c} + \boldsymbol{\beta}_2^t\mathbf{W})$, i.e. if the proportional hazards model holds in the perfectly measured covariates, if $\lambda(t|\mathbf{c}, \mathbf{C}, \mathbf{W}) = \lambda(t|\mathbf{c}, \mathbf{W})$, i.e. measurement error is non-differential and if $\lambda(t|\mathbf{C}, \text{no censorship in}[0,t)) = \lambda(t|\mathbf{C})$, i.e. if there is random censorship conditional on the observed main study data, then

$$\begin{aligned}
\lambda(t|\mathbf{C}, \mathbf{W}) &= \lambda_0(t)\mathrm{E}(\exp(\boldsymbol{\beta}_1^t\mathbf{c} + \boldsymbol{\beta}_2^t\mathbf{W})|\mathbf{C}, \mathbf{W}, T \geqslant t) \\
&= \lambda_0(t)\exp(\boldsymbol{\beta}_2^t\mathbf{W})\mathrm{E}(\exp(\boldsymbol{\beta}_1^t\mathbf{c})|\mathbf{C}, \mathbf{W}, T \geqslant t) \\
&\approx \lambda_0(t)\exp(\boldsymbol{\beta}_2^t\mathbf{W})\mathrm{E}(\exp(\boldsymbol{\beta}_1^t\mathbf{c})|\mathbf{C}, \mathbf{W}),
\end{aligned} \tag{2}$$

where $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are respectively p_1-vector and p_2-vector of regression coefficients corresponding to $\mathbf{c}$ and $\mathbf{W}$, and following Prentice (1982), $T \geqslant t$ can be dropped out when the event is rare.

From (2), we see that the critical quantity is $\mathrm{E}(\exp(\boldsymbol{\beta}_1^t\mathbf{c})|\mathbf{C},\mathbf{W})$. There are two basic ways of dealing with this quantity: exact evaluation or approximation. Exact evaluation requires assuming a model for the full distribution of $(\mathbf{c}|\mathbf{C},\mathbf{W})$. Approximation can be carried out using moment assumptions only. The simplest approximation involves modeling only the conditional mean $\boldsymbol{\mu}_\mathbf{c}(\mathbf{C},\mathbf{W}) = \mathrm{E}(\mathbf{c}|\mathbf{C},\mathbf{W})$, and uses the first-order approximation $\mathrm{E}(\exp(\boldsymbol{\beta}_1^t\mathbf{c})|\mathbf{C},\mathbf{W}) \approx \exp(\boldsymbol{\beta}_1^t\boldsymbol{\mu}_\mathbf{c})$. This approach leads naturally to imputing $\boldsymbol{\mu}_\mathbf{c}(\mathbf{C},\mathbf{W})$ for $\mathbf{c}$ and running a standard Cox analysis. A more sophisticated approximation can be carried out by introducing models for both the conditional mean $\boldsymbol{\mu}_\mathbf{c}(\mathbf{C},\mathbf{W}) = \mathrm{E}(\mathbf{c}|\mathbf{C},\mathbf{W})$ and the conditional variance $\Sigma_\mathbf{c}(\mathbf{C},\mathbf{W}) = \mathrm{Cov}(\mathbf{c}|\mathbf{C},\mathbf{W})$. The approximation is given by

$$\lambda(t|\mathbf{C},\mathbf{W}) \;\approx\; \lambda_0(t)\exp(\boldsymbol{\beta}_2^t\mathbf{W})\exp(\boldsymbol{\beta}_1^t\boldsymbol{\mu}_\mathbf{c} + \frac{1}{2}\boldsymbol{\beta}_1^t\Sigma_\mathbf{c}\boldsymbol{\beta}_1), \tag{3}$$

which is obtained from a second-order Taylor approximation to the cumulant generating function of $(\mathbf{c}|\mathbf{C},\mathbf{W})$. In the special case where $(\mathbf{c}|\mathbf{C},\mathbf{W})$ is multivariate normal, the second order approximation is exact (Prentice (1982)); however, the approximation can be used even in the non-normal case. The first-order approximation is the approach most commonly taken.

Equation (3) allows for a semi-parametric error model $(\mathbf{c}|\mathbf{C},\mathbf{W})$, where only the conditional mean and covariance of $(\mathbf{c}|\mathbf{C},\mathbf{W})$, rather than the whole distribution, needs to be specified. For the ordinary regression calibration method, the multivariate results are similar to those given for the logistic regression model in Rosner et al. (1990). For one-dimensional β without any error-free covariates, when the disease is rare, or β is small, or if the measurement error variance is small and constant, the ordinary regression calibration estimator is given by

$\hat{\beta}_{orc} = \hat{\beta}_{naive}/\hat{\alpha_1}$ and $\hat{\text{Var}}(\hat{\beta}_{orc}) = \frac{1}{\hat{\alpha_1^2}}\hat{\text{Var}}(\hat{\beta}) + \frac{\hat{\beta^2}}{\hat{\alpha_1^4}}\hat{\text{Var}}(\hat{\alpha_1})$ (Spiegelman et al., 1997), where $\hat{\beta}_{naive}$ is the naive Cox regression estimate using the surrogate measure C directly, and $\hat{\alpha_1}$ is obtained in the validation study by fitting the linear regression model given by $\text{E}(c|C) = \alpha_0 + \alpha_1 C$ and $\text{Var}(c|C) = \sigma^2$.

3. Risk set regression calibration for time-varying exposures in a main study/validation study design

The validity of the ordinary regression calibration method depends on the rare disease assumption, i.e. when $\Pr(T \geqslant t) \approx 1$, as shown in (2). Risk set regression calibration is an attempt to improve the estimator by recalibrating within each risk set (Xie et al., 2001). Here, we consider the first order approximation of (2) as

$$\begin{aligned} \lambda(t|\mathbf{C},\mathbf{W}) &= \lambda_0(t)\text{E}(\exp(\boldsymbol{\beta}_1^t\mathbf{c} + \boldsymbol{\beta}_2^t\mathbf{W})|\mathbf{C},\mathbf{W},T \geqslant t) \\ &\approx \lambda_0(t)\exp(\boldsymbol{\beta}_2^t\mathbf{W})\exp(\boldsymbol{\beta}_1^t\text{E}(\mathbf{c}|\mathbf{C},\mathbf{W},T \geqslant t)). \end{aligned} \tag{4}$$

Although $T \geqslant t$ is retained in (4), whenever $\text{Var}(\mathbf{c}|\mathbf{C},\mathbf{W},T \geqslant t)$ and higher order moments are not constants or are not independent of time, an asymptotic bias is expected to be incorporated due to the effect of the higher order moments. Xie et al. (2001) explored analytically the asymptotic bias of the RRC estimate and derived the sandwich variance estimator for the main/reliability study design, in which one or more additional measurements are obtained from a random subsample of study subjects, under the assumption that the classic additive homoscedastic error model is suitable for the data at hand. Since the assumption of classical additive error in a time-invariant exposure is rarely suitable in nutritional and environmental epidemiology, it was necessary to extend the risk set regression calibration method to time-varying exposures in the main study/external validation study design, assuming a more general measurement error model.

With a time-varying point exposure, a main/external validation study design consists of data $\{\mathbf{C}_i(t), \mathbf{W}_i(t), T_i, D_i\}$, $i = 1, \cdots, n_1$ in the main study, and $\{\mathbf{c}_i(t), \mathbf{C}_i(t), \mathbf{W}_i(t), T_i\}$, $i = n_1 + 1, \cdots, n$ in the validation study, where the time-varying surrogate exposure $\mathbf{C}_i(t)$ is measured at a discrete grid of time points and the true exposure $\mathbf{c}_i(t)$ is also measured on certain occasions in the validation study. The gold standard $\mathbf{c}_i(t)$ is usually measured much less frequently than the surrogate $\mathbf{C}_i(t)$.

The exposure variables of interest in these studies are, as noted in the introduction, generally some function of the time-varying point exposures. Denote the function of the time-varying true exposure $\mathbf{c}_i(t)$ as $\mathbf{x}_i(t)$, and the function of the time-varying surrogate exposure $\mathbf{C}_i(t)$ as $\mathbf{X}_i(t)$. Time-varying error-free covariates, and functions of these error-free covariates, are denoted by $\mathbf{Z}_i(t)$. Then, the main/external validation study data take the form $\{\mathbf{X}_i(t), \mathbf{Z}_i(t), T_i, D_i\}$, $i = 1, \cdots, n_1$ for the main study, and $\{\mathbf{x}_i(t), \mathbf{X}_i(t), \mathbf{Z}_i(t), T_i\}$, $i = n_1 + 1, \cdots, n$ for the validation study, and (4) becomes

$$\lambda(t|\mathbf{X}(t), \mathbf{Z}(t)) \approx \lambda_0(t) \exp(\boldsymbol{\beta}_1^t \mathrm{E}(\mathbf{x}(t)|\mathbf{X}(t), \mathbf{Z}(t), T \geqslant t) + \boldsymbol{\beta}_2^t \mathbf{Z}(t)).$$

The basic idea here is that the measurement error model

$$\mathrm{E}(\mathbf{x}_i(t)|\mathbf{X}_i(t), \mathbf{Z}_i(t)) = \boldsymbol{\alpha}_0(t) + \boldsymbol{\alpha}_1(t)\mathbf{X}_i(t) + \boldsymbol{\alpha}_2(t)\mathbf{Z}_i(t) \tag{5}$$

is re-estimated from the validation study at each main study failure time, and the true exposure is re-estimated for everyone at risk. Then, $\hat{\beta}_{RRC}$ will solve

$$\sum_{i=1}^{n_1} \int_0^{t^*} \left\{ \begin{pmatrix} \hat{\mathbf{x}}_i(\hat{\boldsymbol{\psi}}, t) \\ \mathbf{Z}_i(t) \end{pmatrix} - \frac{\mathbf{S}^{(1)}(\boldsymbol{\beta}, \hat{\boldsymbol{\psi}}, t)}{S^{(0)}(\boldsymbol{\beta}, \hat{\boldsymbol{\psi}}, t)} \right\} N_i(\, dt) = 0 \tag{6}$$

where

$$S^{(0)}(\boldsymbol{\beta}, \hat{\boldsymbol{\psi}}, t) = {n_1}^{-1} \sum_{i=1}^{n_1} Y_m(i, t) \exp\{\boldsymbol{\beta}_1^t \hat{\mathbf{x}}_i(\hat{\boldsymbol{\psi}}, t) + \boldsymbol{\beta}_2^t \mathbf{Z}_i(t)\},$$

and

$$\mathbf{S}^{(1)}(\boldsymbol{\beta}, \hat{\boldsymbol{\psi}}, t) = {n_1}^{-1} \sum_{i=1}^{n_1} Y_m(i,t) \begin{pmatrix} \hat{\mathbf{x}}_i(\hat{\boldsymbol{\psi}}, t) \\ \mathbf{Z}_i(t) \end{pmatrix} \exp\{\boldsymbol{\beta}_1^t \hat{\mathbf{x}}_i(\hat{\boldsymbol{\psi}}, t) + \boldsymbol{\beta}_2^t \mathbf{Z}_i(t)\}.$$

Here, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$, $N_i(t) = I(T_i \leqslant t, D_i = 1)$ is the counting process corresponding to T_i. $N_i(dt) = 1$ if the subject i fails at the failure time t, $N_i(dt) = 0$ otherwise. $Y_m(i, t_l) = I(T_i \geqslant t_l)$ is the risk process indicator in the main study, the subscript 'm' indicates the main study, the subscript 'v' indicates the validation study. $\dim(\boldsymbol{\alpha}_0(t_l)) = p_1$, $\dim(\boldsymbol{\alpha}_1(t_l)) = p_1 \times p_1$, $\dim(\boldsymbol{\alpha}_2(t_l)) = p_1 \times p_2$, $\boldsymbol{\psi}(t_l) = (\boldsymbol{\alpha}_0(t_l), \boldsymbol{\alpha}_1(t_l), \boldsymbol{\alpha}_2(t_l))$, $l = 1, \cdots, r$, and r is the number of unique failure time in the main study. Next, we explain how to obtain estimates of $\boldsymbol{\alpha}_0(t_l)$, $\boldsymbol{\alpha}_1(t_l)$, $\boldsymbol{\alpha}_2(t_l)$:

1. Order the unique failure times that occur in the main study as $t_1 < t_2 < \cdots < t_r$. Find the r risk sets $R_m(t_l)$, $l = 1, 2, \cdots, r$, where $R_m(t_l)$ is the set of individuals in the main study who are alive and uncensored at a time just prior to t_l. Generate the risk process indicator $Y_m(i, t_l)$ so that $Y_m(i, t_l) = 1$ if $i \in R_m(t_l)$, and $Y_m(i, t_l) = 0$ otherwise.
2. Find the r risk sets $R_v(t_l)$, $l = 1, 2, \cdots, r$, in the validation study. Generate the risk process indicator $Y_v(i, t_l) = I(T_i \geqslant t_l)$, so that $Y_v(i, t_l) = 1$ if $i \in R_v(t_l)$, and $Y_v(i, t_l) = 0$ otherwise. For any given t, let $\mathbf{X}_i^*(t)$, $\mathbf{x}_i^*(t)$ and $\mathbf{Z}_i^*(t)$ be the most recent observed values of $\mathbf{X}_i(t)$, $\mathbf{x}_i(t)$ and $\mathbf{Z}_i(t)$ prior to t. We then run, for each failure time t_l, a regression of $\mathbf{x}_i^*(t_l)$ on $\mathbf{X}_i^*(t_l)$ and $\mathbf{Z}_i^*(t_l)$ on the subjects in $R_v(t_l)$, obtaining $\hat{\boldsymbol{\alpha}}_0(t_l)$, $\hat{\boldsymbol{\alpha}}_1(t_l)$, $\hat{\boldsymbol{\alpha}}_2(t_l)$.
3. Estimate $\hat{\mathbf{x}}_i(t_l) = [\hat{\boldsymbol{\alpha}}_0(t_l) + \hat{\boldsymbol{\alpha}}_1(t_l)\mathbf{X}_i^*(t_l) + \hat{\boldsymbol{\alpha}}_2(t_l)\mathbf{Z}_i^*(t_l)] \cdot Y_m(i, t_l)$ for each subject i in the risk set $R_m(t_l)$ in the main study, $l = 1, 2, \cdots, r$.
4. Fit the "naive" cox model on $(\hat{\mathbf{x}}_i(t_l), \mathbf{Z}_i(t_l), Y_m(i, t_l), T_i, D_i)$ to get the risk set regression calibration estimator $\hat{\boldsymbol{\beta}}_{RRC}$.

Non-differential measurement error and random censorship are also required here to ensure validity of estimation and inference, as in the original regression calibration estimate

discussed in Section 2. In addition, we assume that the measurement error model (5) that holds in the validation study is applicable to the main study as well. This assumption is the transportability assumption discussed by Carroll et al. (2006).

4. $\mathbf{Var}(\hat{\boldsymbol{\beta}}_{rrc})$ for the main/ external validation study

We write the score equation (6) as $\sum_{i=1}^{n_1} \mathbf{U}_{\boldsymbol{\beta} i}(\boldsymbol{\beta}|\hat{\boldsymbol{\psi}}) = \mathbf{0}$, where

$$\mathbf{U}_{\boldsymbol{\beta} i}(\boldsymbol{\beta}|\hat{\boldsymbol{\psi}}) = D_i \left\{ \begin{pmatrix} \hat{\mathbf{x}}_i(T_i) \\ \mathbf{Z}_i(T_i) \end{pmatrix} - \frac{\mathbf{S}^{(1)}(\boldsymbol{\beta}, \hat{\boldsymbol{\psi}}, T_i)}{S^{(0)}(\boldsymbol{\beta}, \hat{\boldsymbol{\psi}}, T_i)} \right\}, \tag{7}$$

with $\hat{\boldsymbol{\psi}} = (\hat{\boldsymbol{\psi}}(t_1), \hat{\boldsymbol{\psi}}(t_2), \cdots, \hat{\boldsymbol{\psi}}(t_r))$, $\dim(\hat{\boldsymbol{\psi}}) = p_1 \times (p_1+p_2+1)r$ and $\dim(\mathbf{U}_{\boldsymbol{\beta} i}) = (p_1+p_2)\times 1$. Define $\hat{\mathbf{U}}^*(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\psi}}) = \sum_{i=1}^{n_1} \frac{\partial \mathbf{U}_{\boldsymbol{\beta} i}(\boldsymbol{\beta}|\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}|_{(\boldsymbol{\beta},\boldsymbol{\psi})=(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\psi}})}$, then $\dim(\hat{\mathbf{U}}^*) = (p_1+p_2) \times p_1(p_1+p_2+1)r$. Denote the covariance matrix of $\hat{\boldsymbol{\psi}}$ as $\mathbf{V}_{\hat{\boldsymbol{\psi}}}$, so $\dim(\mathbf{V}_{\hat{\boldsymbol{\psi}}}) = p_1(p_1+p_2+1)r \times p_1(p_1+p_2+1)r$.

In the validation study, $\hat{\boldsymbol{\psi}}$ solves

$$\sum_{i=1}^{n_2} \mathbf{U}_{\boldsymbol{\psi} i,j}(\boldsymbol{\psi}) = \mathbf{0}, \tag{8}$$

where $\mathbf{U}_{\boldsymbol{\psi} i,j}(\boldsymbol{\psi}) = (\mathbf{U}^t_{\boldsymbol{\alpha}_0 i,j}(\boldsymbol{\psi}), \mathbf{U}^t_{\boldsymbol{\alpha}_1 i,j}(\boldsymbol{\psi}), \mathbf{U}^t_{\boldsymbol{\alpha}_2 i,j}(\boldsymbol{\psi}))$ is defined as follows

$$\begin{aligned} \mathbf{U}_{\boldsymbol{\alpha}_0 i,j}(\boldsymbol{\psi}) &= Y_v(i,t_j)[\mathbf{x}_i - \boldsymbol{\alpha}_0(t_j) - \boldsymbol{\alpha}_1(t_j)\mathbf{X}_i - \boldsymbol{\alpha}_2(t_j)\mathbf{Z}_i]^t \\ \mathbf{U}_{\boldsymbol{\alpha}_1 i,j}(\boldsymbol{\psi}) &= Y_v(i,t_j)\mathbf{X}_i[\mathbf{x}_i - \boldsymbol{\alpha}_0(t_j) - \boldsymbol{\alpha}_1(t_j)\mathbf{X}_i - \boldsymbol{\alpha}_2(t_j)\mathbf{Z}_i]^t \\ \mathbf{U}_{\boldsymbol{\alpha}_2 i,j}(\boldsymbol{\psi}) &= Y_v(i,t_j)\mathbf{Z}_i[\mathbf{x}_i - \boldsymbol{\alpha}_0(t_j) - \boldsymbol{\alpha}_1(t_j)\mathbf{X}_i - \boldsymbol{\alpha}_2(t_j)\mathbf{Z}_i]^t \end{aligned} \tag{9}$$

for $j = 1, \cdots, r$, $i = 1, \cdots, n_2$, $\dim(\mathbf{U}_{\boldsymbol{\alpha}_0 i,j}) = 1 \times p_1$, $\dim(\mathbf{U}_{\boldsymbol{\alpha}_1 i,j}) = p_1 \times p_1$, $\dim(\mathbf{U}_{\boldsymbol{\alpha}_2 i,j}) = p_2 \times p_1$, $\dim(\mathbf{U}_{\boldsymbol{\psi} i,j}(\boldsymbol{\psi})) = p_1 \times (p_1+p_2+1)$, $\mathbf{U}_{\boldsymbol{\psi} i} = (\mathbf{U}_{\boldsymbol{\psi} i,1}, \mathbf{U}_{\boldsymbol{\psi} i,2}, \cdots, \mathbf{U}_{\boldsymbol{\psi} i,r})$ and $\dim(\mathbf{U}_{\boldsymbol{\psi} i}) = p_1 \times (p_1+p_2+1)r$.

Then, as shown in Appendix B, $\mathrm{Cov}(\hat{\boldsymbol{\psi}})$ can be estimated by $\frac{1}{n_2}\hat{\mathbf{V}}_{\hat{\boldsymbol{\psi}}}$, with $\hat{\mathbf{V}}_{\hat{\boldsymbol{\psi}}}$ constructed as

$$\hat{\mathbf{V}}_{\hat{\boldsymbol{\psi}}} = \left[\frac{1}{n_2}\sum_{i=1}^{n_2} \frac{\partial \mathbf{U}_{\boldsymbol{\psi} i}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}\right]^{-1} \left[\frac{1}{n_2}\sum_{i=1}^{n_2} \mathbf{U}_{\boldsymbol{\psi} i}(\boldsymbol{\psi}) \otimes \mathbf{U}^T_{\boldsymbol{\psi} i}(\boldsymbol{\psi})\right] \left[\frac{1}{n_2}\sum_{i=1}^{n_2} \frac{\partial \mathbf{U}_{\boldsymbol{\psi} i}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}\right]^{-1} \Big|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}}. \tag{10}$$

Following the derivation in Appendix B, we have

$$\hat{\text{Var}}(\hat{\boldsymbol{\beta}}_{rrc}) \;=\; \frac{1}{n_1}\hat{\mathbf{I}}_{\boldsymbol{\beta}}^{-1}\hat{\mathbf{H}}_{\boldsymbol{\beta},\boldsymbol{\psi}}\hat{\mathbf{I}}_{\boldsymbol{\beta}}^{-1} \tag{11}$$

where

$$\hat{\boldsymbol{I}}_{\boldsymbol{\beta}}^{-1} = \left[\frac{1}{n_1}\sum_{i=1}^{n_1}\frac{\partial \mathbf{U}_{\boldsymbol{\beta}i}(\boldsymbol{\beta}|\hat{\boldsymbol{\psi}})}{\partial \boldsymbol{\beta}}\right]^{-1}\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}, \tag{12}$$

$$\hat{\mathbf{H}}_{\boldsymbol{\beta},\boldsymbol{\psi}} = \hat{\mathbf{H}}_{\boldsymbol{\beta}} + \frac{1}{n_1 n_2}\hat{\mathbf{U}}^*(\boldsymbol{\beta},\boldsymbol{\psi})\hat{\mathbf{V}}_{\boldsymbol{\psi}}\hat{\mathbf{U}}^*(\boldsymbol{\beta},\boldsymbol{\psi})^T,$$

$$\hat{\mathbf{H}}_{\boldsymbol{\beta}} = \frac{1}{n_1}\sum_{i=1}^{n_1}\tilde{\mathbf{U}}_{\boldsymbol{\beta}i}(\boldsymbol{\beta},\hat{\boldsymbol{\psi}})\tilde{\mathbf{U}}_{\boldsymbol{\beta}i}^T(\boldsymbol{\beta},\hat{\boldsymbol{\psi}})|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}, \tag{13}$$

with

$$\tilde{\mathbf{U}}_{\boldsymbol{\beta}i}(\boldsymbol{\beta},\hat{\boldsymbol{\psi}}) = \mathbf{U}_{\boldsymbol{\beta}i}(\boldsymbol{\beta}|\hat{\boldsymbol{\psi}}) - \sum_{j=1}^{n_1}\frac{D_j Y_m(i,T_j)\exp(\boldsymbol{\beta}_1^t\hat{\mathbf{x}}_i(T_j)+\boldsymbol{\beta}_2^t\mathbf{Z}_i(T_j))}{n_1 S^{(0)}(\boldsymbol{\beta},\hat{\boldsymbol{\psi}},T_j)}$$
$$\cdot\left\{\begin{pmatrix}\hat{\mathbf{x}}_i(\hat{\boldsymbol{\psi}},T_j)\\ \mathbf{Z}_i(T_j)\end{pmatrix} - \frac{\mathbf{S}^{(1)}(\boldsymbol{\beta},\hat{\boldsymbol{\psi}},T_j)}{S^{(0)}(\boldsymbol{\beta},\hat{\boldsymbol{\psi}},T_j)}\right\}.$$

The asymptotic distribution theory proven in Appendix B follows arguments in Andersen and Gill (1982) and Lin and Wei (1989); the reader is referred to Appendix B for further details.

5. A simulation study under the main study/external validation study

We report finite-sample simulation results for our proposed method under the main study / external validation study design, following the algorithm in Sections 3 and 4. We consider a single error-prone covariate c with surrogate C. All simulation results are based on 1000 replications.

We consider two event rate scenarios: rare disease and common disease. Motivated by our real data set, the Health Professional Follow-up Study, we set the number of events to be around 500. For the rare disease situation, we set the cumulative event rate at 1%, and thus

the main study sample size is set at $n_1 = 50,000$. For the common disease situation, we set the cumulative event rate at 50%, and thus the main study sample size is set at n_1=1000. The cases we considered for the validation study size were $n_2 = 150$ and $n_2 = 1,000$. A validation sample size of $n_2 = 150$ is very common in nutritional epidemiology studies. The case of $n_2 = 500$ is less common, but does arise in some applications, particularly when two or more validation studies can be combined, as in the example in Section 6.

We take the baseline hazard function to be of the Weibull form $\lambda_0(t) = \theta\nu(\nu t)^{\theta-1}$, with $\theta = 6$, which is typical of many epithelial cancers (Armitage and Doll (1961); Breslow and Day (1993)). Censoring is assumed exponential with a rate of 1% per year. The maximum follow-up time is taken to be t^* =50 years. The parameter ν is set to achieve the specified cumulative event rate.

In preliminary simulations, we examined the case where the covariate is time-invariant, and compared the ORC method to the RRC method. These results are summarized in Appendix A.1. The parameter which describes the extent of measurement error, $\rho = \mathrm{Corr}(c, C)$ was varied across the values 0.3, 0.6, 0.9. The key parameter in the simulation comparisons is $\eta = \beta^2\mathrm{Var}(c|C) = (1-\rho^2)\beta^2\sigma^2$, where β is the regression coefficient and σ^2 is the variance of the true exposure. When η is small, the ORC method was adequate, but when η is large, the RRC method was clearly superior. These findings are consistent with those of Xie et al. (2001).

We turn now to the time-varying exposure situation. In line with the motivating example in Section 6, we worked with the cumulatively updated average exposure (Hu et al., 1999). The true and surrogate exposures, $\mathbf{x}(t)$ and $\mathbf{X}(t)$, were defined as follows:

$$\mathbf{x}(t) = \frac{1}{t_k - t_0}\sum_{m=1}^{k}\mathbf{c}(t_{m-1})(t_m - t_{m-1}), \quad \mathbf{X}(t) = \frac{1}{t_k - t_0}\sum_{m=1}^{k}\mathbf{C}(t_{m-1})(t_m - t_{m-1}), \qquad (14)$$

for $t \in [t_k, t_{k+1})$ with $1 \leqslant k \leqslant p$, where the set $\{t_0, t_1, t_2, \cdots, t_p\}$ are the times at which $\mathbf{c}_i(t)$ is measured. $\mathbf{C}_i(t)$ can be measured on the same time scale, or, as is typically the case in

applications, $\mathbf{C}_i(t)$ is measured on a much finer time scale. In the simulation study below, we used the same time scale for $\mathbf{c}_i(t)$ and $\mathbf{C}_i(t)$.

The true and surrogate cumulatively updated average exposures were generated as follows:

1. The true exposure $\boldsymbol{c}_i \sim MVN(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$, where $\boldsymbol{c}_i$ is a $p-$vector with p as the number of the observation time points, $\boldsymbol{\mu}_c$ is the mean vector and $\boldsymbol{\Sigma}_c$ is a covariance matrix. Without loss of generality, we consider a simple case with $\boldsymbol{\mu}_c = \mathbf{0}$ and $\boldsymbol{\Sigma}_c$ such that $\Sigma_c(j,k) = 1$ if $j = k$ and $\Sigma_c(j,k) = \rho_I^{|j-k|^\tau}$ if $j \neq k$, with $\tau = 0$ or 1. When $\tau = 0$, a compound symmetry covariance structure is obtained, and the intra-class correlation $\rho_{I_{CS}}$ was set at 0.3, 0.6, 0.9. When $\tau = 1$, an AR(1) structure is obtained. To put these two covariance scenarios on an equal footing, we set the average correlation of $\rho_{I_{AR}}$ over $[0, t^*]$ equal to $\rho_{I_{CS}}$, i.e.,

$$\frac{1}{t^*}\int_0^{t^*} (\rho_{I_{AR}})^t dt = \rho_{I_{CS}}. \tag{15}$$

Solving this equation, we obtained the corresponding values of $\rho_{I_{AR}}$ as 0.938, 0.978, 0.996.

2. $C_{ij} = c_{ij} + e_{ij}$, $e_{ij} \sim N(0, \Delta)$, with the measurement error variance Δ given by $\Delta = \frac{1}{\rho^2} - 1$, where ρ is the correlation between C_{ij} and c_{ij} and varied as 0.3, 0.6, 0.9.

3. A cumulative average exposure $x(t)$ and $X(t)$ was then generated by

$$x_i(t_k) = \frac{1}{t_k - t_0}\sum_{m=1}^{k} c_i(t_{m-1})(t_m - t_{m-1}), \quad X_i(t_k) = \frac{1}{t_k - t_0}\sum_{m=1}^{k} C_i(t_{m-1})(t_m - t_{m-1}),$$

for $1 \leqslant k \leqslant p$, over the time points $\{t_0, t_1, t_2, \cdots, t_p\}$. In applications, time could be expressed in terms of participant's age, time on the study, calendar year, or in some other appropriate manner. In our simulations, for simplicity, we set $t_j = 5 * j$, for $j = 0, 1, \cdots, 10$. Thus, there were 10 exposure measurements in total over the study period of $t^* = 50$ years. This scheme was patterned after the measurement schedule in our motivating example, HPFS, and the other Harvard cohort studies.

The survival data were generated according to the hazard model $\lambda(t|x(t)) = \lambda_0(t)\exp(\beta x(t))$, where $x(t)$ is as defined in (14). Appendix A.2 presents details of the data generation. Since survival is usually measured on a finer time scale than that defined by the exposure measurement schedule, we generated the survival time on a finer time scale, based on $t_j = j$, for $j = 0, 1, \cdots, 50$.

Table 1 presents the results for the compound symmetry structure. The results for the AR(1) structure, which were similar, are presented in Appendix A.3. Overall, the RRC estimator performed very well in this time-varying cumulatively updated average exposure setting, with the good performance being robust across different levels of the autocorrelation in the true exposure process and the different correlation structures. The bias was very small for all scenarios and became even smaller as the autocorrelation became higher. An increased bias was seen with the AR(1) correlation structure when the autocorrelation was low, consistent with the fact that within-person variability increases more quickly with time for the AR(1) structure than the CS structure. However, the worst bias seen was only 5%. The coverage probability was nearly accurate in all cases considered. Increasing the frequency of exposure measurements improved the results even more (data not shown), although the bias was already minimal with the exposure frequencies investigated here.

[Table 1 about here.]

6. Motivating example

We illustrated our method in the Health Professionals' Follow-Up Study (HPFS) of the relationship between the total calcium intake and risk of fatal prostate cancer (Giovannucci et al., 2006). HPFS is an ongoing prospective cohort study of cancer and heart disease among 51529 U.S. male health professionals who responded to a mailed baseline questionnaire in 1986, asking about demographics, family history of disease, diet, smoking, physical activity,

medications and other lifestyle factors. Every two years, study participants receive questionnaires to update health status information and potential risk factors. Total vitamin E intake was the only important confounder in this study. The food frequency questionnaire (FFQ) were administered in 1986, 1990, 1994, 1998 and 2002 to assess dietary intake, including total calcium and vitamin E. After excluding men with a history of cancer at baseline, or who did not adequately complete the 1986 dietary questionnaire, there were 390703 person-year observed in the main study with 357 fatal prostate cancer cases among 47760 subjects between 1986 to 2008. In our analysis, we used age as the time scale, as is more suitable for epidemiologic studies of chronic disease (Korn, Graubard, and Midthune, 1997), hence a left-truncated analysis is implied here.

The FFQ measures dietary intake with some degree of error and more reliable information can be obtained from a diet record(DR) (Willett, 1998). In the HPFS validation study, 2 weeks of weighed diet records (DR) were observed in 127 person-years among 127 study participants. To increase the validation study sample size, we included another dietary validation study, the Eating at America's Table Study (EATS) to our analysis (Subar et al., 2001). EATS is a study that was designed to validate the Diet History Questionnaire (DHQ), a food frequency questionnaire (FFQ) similar to the one used in HPFS. In EATS, the exposures was validated by four telephone-administered 24-hour dietary recalls. The left half of Table 2 compares the basic characteristics of these two validation studies, which are very similar. To ascertain the validity of the transportability assumption of EATS to HPFS, we ran a regression analysis of DR on FFQ for each study, adjusting for age in 5-year age groups. The slope for total calcium intake was 0.445 in EATS and 0.371 in HPFS. For total vitamin E, the slope was 0.818 in EATS and 0.762 in HPFS. Because they were so similar, we accepted the reasonableness of the transportability assumption here. The rightmost major

column of Table 2 shows the basic characteristics of the HPFS main study, so they can be compared to the characteristics of the two validation studies.

We empirically investigated the suitability of the transportability assumption with further analysis. We defined a binary indicator $Study_{ind}$ with value 1 if the study was HPFS and 0 if the study was EATS, and then ran the regression model

$$Calc_{dr} = \gamma_0 + \gamma_1 Calc_{ffq} + \gamma_2 Calc_{ffq} * Study_{ind}, \tag{16}$$

and similarly for vitamin E. To test the hypothesis that there is no between-studies variation in the slope, i.e. $\gamma_2 = 0$ by age group, we fit regression models by 5-year age groups to the combined validation study, using model (16). We found that the null hypothesis was accepted for most age groups, except in age group 56 - 60, the $p-$value for the total calcium intake by study interaction was less than 0.001, and in age group older than 71, the $p-$value for the total vitamin E intake by study interaction was less than 0.03. So in our analysis, when we considered both total calcium intake and total vitamin E intake as error-prone exposures, we excluded the EATS study participants from the risk sets in those ages. When we considered only the total calcium intake to be measured with error, we kept the EATS study participants in the risk set of age 71 and older, but we excluded the EATS study subjects from the age of 56 - 60 risk sets.

Table 3 gives the results of the RRC method for time-varying exposures developed in this paper compared to the naive Cox approach. The relative risks are given in units of 1838 mg/day for total calcium intake to facilitate comparison of the result reported in the original publication of these data (Giovannucci et al., 2006), where 1838 mg/day was the difference between the median of the top quintile and the median of bottom quintile. We performed the analysis using only the HPFS validation study and using both HPFS and EATS validation studies. Because the HPFS validation study had 127 person-year observations for 127 subjects, we grouped the risk sets by 4-year age interval to stabilize

the estimates. This was not necessary when we included both validation studies. The results show that with more data in the validation study, more stable and reasonable estimates are obtained. From the results including both HPFS and EATS validation studies, we can see that the under-estimation of the effect of calcium intake on fatal prostate cancer from the naive Cox approach was corrected by the RRC estimate. Total calcium intake had a significant positive association with fatal prostate cancer and the total vitamin E intake was weakly inversely associated, perhaps not associated at all.

[Table 2 about here.]

[Table 3 about here.]

7. Discussion

In this paper, we have considered the Cox survival regression model with time-dependent covariates subject to measurement error. We derived a bias-corrected point and interval estimate of the relative risk using a RRC approach. We emphasized the main study/external validation study design, which arises commonly in nutritional epidemiology, but has been given less attention in the literature than other designs, particularly in the setting of time-varying covariates. We focused on cumulatively updated total or average exposure, which is often of interest in nutritional and environmental epidemiology, but the approach could also be applied to analysis of time-varying point exposures. In addition, the approach extends in a straightforward way to cover left truncation. A FORTRAN program for the method is available at `http://www.hsph.harvard.edu/faculty/spiegelman/rrc_timevarying_method.f`. Simulation studies in the setting of cumulatively updated average exposure showed that the method performs very well.

Clearly, the effectiveness of the method depends on the size of the validation study. The analysis of the HPFS data in Section 6 illustrates this point. When we used only the data

from the 127 participants in the HPFS validation study, and with vitamin E assumed to be measured without error, a very wide confidence interval for the effect of calcium intake was obtained, leading to results that were not very meaningful. When we included validation data data from the 573 men in the EATS study, the results were much more reasonable, and significant p-values for both calcium intake and vitamin E were obtained, even when both these nutrients were regarded as subject to with error.

To date, most validation studies available in nutritional epidemiology have only one measurement per subject. Only the Nurses Health Study has validation data repeatedly assessed over time, and this is among a small number of subjects. The limited validation data makes it challenging to implement measurement error correction for time-varying exposures. However, we understand that new validation studies are underway, including one within our own group at the Nurses Health Study, involving a much larger number of subjects and repeated exposure assessments over time. Such expanded validation studies will provide a firmer basis of measurement error correction, and the method we have developed here provides a practical and reliable way for carrying out such correction.

In summary, measurement error in time-varying covariates, including those that are functions of a series of exposure measurements available up to a failure time, is an extremely common problem in nutritional and environmental epidemiology. The method developed in this paper provides a mechanism for handling this problem that is well-suited to these applications.

References

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting process: a large sample study. *The Annals of Statistics* **10,** 1100 – 1120.

Armitage, P. and Doll, R. (1961). Stochastic models for carcinogenesis. In *Neyman,*

J., ed., Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, pages 19–38. University of California Press, Berkeley.

Brauer, M., Hoek, G., van Vliet, P., Meliefste, K., Fischer, P., Gehring, U., Heinrich, J., Cyrys, J., Bellander, T., Lewne, M., and Brunekreef, B. (2003). Estimating long-term average particulate air pollution concentrations: Application of traffic indicators and geographic information systems. *Epidemiology* **14,** 228–239.

Breslow, N. E. and Day, N. E. (1993). *Statistical Methods in Cancer Research.* World Health Organization.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. (2006). *Measurement error in nonlinear models: A Modern Perspective, Second Edition.* Chapman & Hall.

Cox, D. (1972). Regression models and life tables(with discussion). *Journal of the Royal Statistical Society, Series B* **34,** 187–220.

Giovannucci, E., Liu, Y., Stampfer, M. J., and Willett, W. C. (2006). A prospective study of calcium intake and incident and fatal prostate cancer. *Cancer Epidemiology Biomarkers & Prevention* **15,** 203–210.

Hu, F. B., Stampfer, M. J., Rimm, E., Ascherio, A., Rosner, B. A., Spiegelman, D., and Willett, W. C. (1999). Dietary fat and coronary heart disease: a comparison of approaches to adjusting for total energy intake and modeling repeated dietary measurements. *American Journal of Epidemiology* **149,** 531–540.

Huang, Y. and Wang, C. Y. (2000). Cox regression with accurate covariates unascertainable: a nonparametric-correction approach**45,** 1209–1219.

Hunter, D. J., Spiegelman, D., Adami, H. O., Beeson, L., van den Brandt, P. A., Folsom, A. R., Fraser, G. E., Goldbohm, R. A., Graham, S., Howe, G. R., Kushi, L. H., Marshall, J. R., McDermott, A., Miller, A. B., Speizer, F. E., Wolk, A., Yaun, S. S., and Willett, W. (1996). Cohort studies of fat intake and risk of breast cancer: a pooled analysis. *New*

England Journal of Medicine **334,** 356–361.

Korn, E. L., Graubard, B. I., and Midthune, D. (1997). Time-to-event analysis of longitudinal follow-up of a survey: Choice of the time-scale. *American Journal of Epidemiology* **145,** 72 – 80.

Lin, D. and Wei, L. (1989). The robust inference for the cox proportional hazards model. *Journal of the American Statistical Association* **84,** 1074–1078.

Prentice, R. (1982). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika* **69,** 331–342.

Raboud, J., Reid, N., Coates, R., and Farewell, V. (1993). Estimating risks of progressing to aids when covariates are measured with error. *Journal of the Royal Statistical Society, Series A* **156,** 393–406.

Rosner, B., Spiegelman, D., and Willett, W. C. (1990). Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error. *American Journal of Epidemiology* **132,** 734–745.

Spiegelman, D., McDermott, A., and Rosner, B. (1997). Regression calibration method for correcting measurement-error bias in nutritional epidemiology. *American Journal of Clinical Nutrition* **65(suppl),** 1179S–1186S.

Steenland, K. and Savitz, D. A. (1997). *Topics in Environmental Epidemiology.* New York: Oxford University Press.

Subar, A., Thompson, F., Kipnis, V., Midthune, D., Hurwitz, P., McNutt, S., McIntosh, A., and Rosenfeld, S. (2001). Comparative validation of the block, willett, and national cancer institute food frequency questionnaires: The eating at america's table study. *American Journal of Epidemiology* **154,** 1089–1099.

Thomas, D. C. (2009). *Statistical Methods in Environmental Epidemiology.* New York: Oxford University Press.

Tsiatis, A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica* **14,** 809–834.

Willett, W. C. (1998). *Nutritional Epidemiology (2nd Ed).* Oxford University Press: New York.

Willett, W. C., Sampson, L., Stampfer, M. J., Rosner, B. A., Bain, C., Witschi, J., Hennekens, C. H., and Speizer, F. E. (1985). Reproducibility and validity of a semiquantitative food frequency questionnaire. *American Journal of Epidemiology* **122,** 51–65.

Xie, S. X., Wang, C., and Prentice, R. (2001). A risk set calibration method for failure time regression by using a covariate reliability sample. *Journal of the Royal Statistical Society, Series B* **63,** 855–870.

Zanobetti, A., Wand, M. P., Schwartz, J., and Ryan, L. M. (2000). Generalized additive distributed lag models: quantifying mortality displacement. *Biostatistics* **1,** 279–292.

Zucker, D. (2005). A pseudo partial likelihood method for semi-parametric survival regression with covariate errors. *Journal of the American Statistical Association* **100,** 1264–1277.

Appendices

A. Some details for the simulation study in Section 5

A.1 *Preliminary simulations for time-invariant exposures*

We considered a conditional normal error model with a time-invariant covariate, with key parameters motivated by the Health Professionals' Follow-up Study (HPFS) as considered in Section 6. In this model, we first generated the true exposure $c \sim N(\mathrm{E}(c), \mathrm{Var}(c))$ with $\mathrm{E}(c) = 0.45$, $\mathrm{Var}(c) = 0.0225$ as in HPFS. The surrogate exposure C has $\mathrm{E}(C) = 0.5$, $\mathrm{Var}(C) = 0.04$. Define $\omega = \mathrm{Var}(C)/\mathrm{Var}(c)$. For each c, we generated the surrogate exposure C from the conditional distribution $C|c$, which also had a normal distribution with conditional

mean $\mathrm{E}(C|c) = \alpha + \xi c$ and variance $\mathrm{Var}(C|c) = \omega(1-\rho^2)\mathrm{Var}(c)$, where $\rho = \mathrm{Corr}(c, C)$, which we allowed to vary as 0.3, 0.6, 0.9, $\xi = \rho\sqrt{\omega}$ and $\alpha = \mathrm{E}(C) - \xi\mathrm{E}(c)$.

The survival time T^0 was generated by $T^0 = \frac{1}{\nu}(-e^{-\beta c}\log(1-U_1))^{1/\theta}$ with $U_1 \sim U(0,1)$. Then, the follow-up time, $T = \min(T^0, V, t^*)$, for $t^* = 50$ and V is the censoring time assuming to be exponential with a rate of 1% per year. And, the event indicator, $D = I(T^0 \leqslant \min(V, t^*))$.

The simulation results are given in Table 4. We found equally good performance of the ORC and RRC methods with $\mathrm{Var}(c) = 0.0225$. When we increased $\mathrm{Var}(c)$ to be greater than 1, for example, as shown in lower part of Table 4, with the means chosen as previously, but with $\mathrm{Var}(c) = 1.0$ and $\mathrm{Var}(C) = 2.0$, then the results indicated a clear advantage of the RRC method over the ORC method, especially in the common disease situation. Additional simulations demonstrated that this advantage became even greater when $\mathrm{Var}(c)$ got even bigger (data not shown).

Figure 1 shows the percent change in the regression slope $\hat{\alpha}_1(t)$ as a function of the failure time t, where the percent change of $\hat{\alpha}_1(t)$ is with respect to the value of $\hat{\alpha}_1$ from the ORC method, and is defined as $100 * [\hat{\alpha}_1(t) - \hat{\alpha}_1]/\hat{\alpha}_1$. We fitted a lowess smoother to the data from 1000 simulations. We can see from Figure 1 that, with a relatively big variance, i.e. $\mathrm{Var}(c) = 1$ in the conditional normal error model, there was a big change of $\hat{\alpha}_1(t)$ with respect to the baseline value of $\hat{\alpha}_1$ estimated by ORC when the disease was common, while the change was much smaller when the disease was rare. However, with a small variance, i.e. $\mathrm{Var}(c) = 0.0225$, the changes in $\hat{\alpha}_1(t)$ over time were both very small no matter whether the disease was common or rare. This exactly explained why the RRC estimates were superior in the scenario with big $\mathrm{Var}(c)$, especially in the common disease situation, and agreed with the results presented in Table 4.

[Table 4 about here.]

[Figure 1 about here.]

A.2 *Simulation of survival data for time-varying exposures*

The following is the way to generate the survival time T^0 for cumulatively updated average exposure $x(t)$.

The cumulative incidence function for T^0 was

$$F(t|x(t)) = 1 - \exp(-\int_0^t \lambda(s|x(s))\,ds) = 1 - \exp(-\theta\nu^\theta \int_0^t s^{\theta-1}\exp(\beta x(s))\,ds) \qquad \text{(A.1)}$$

If $t_k \leqslant t < t_{k+1}$ for some integer k, we next derived the cumulative incidence function for the cumulatively updated average exposure, $x(t)$, which is

$$\begin{aligned} F(t|x(t)) &= 1 - \exp\left\{-\theta\nu^\theta\left(\sum_{i=0}^{k-1}\int_{t_i}^{t_{i+1}} s^{\theta-1}\exp(\beta x(s))\,ds + \int_{t_k}^{t} s^{\theta-1}\exp(\beta x(s))\,ds\right)\right\} \\ &= 1 - \exp\left\{-\theta\nu^\theta\left(\sum_{i=0}^{k-1}\exp(\beta x(t_i))\int_{t_i}^{t_{i+1}} s^{\theta-1}\,ds + \exp(\beta x(t_k))\int_{t_k}^{t} s^{\theta-1}\,ds\right)\right\} \\ &= 1 - \exp\left\{-\nu^\theta\left(\sum_{i=0}^{k-1}\exp(\beta x(t_i))(t_{i+1}^\theta - t_i^\theta) + \exp(\beta x(t_k))(t^\theta - t_k^\theta)\right)\right\} \qquad \text{(A.2)} \end{aligned}$$

with $t_0 = 0$, $x(0) = 0$. For each subject i, we generated the censoring time V_i in the same way as in Appendix A.1. Then, for each subject i, we calculated F_{ij} using (A.2) as

$$F_{ij} = 1 - \exp\left\{-\nu^\theta\left(\sum_{u=0}^{j-1}\exp(\beta x_i(t_u))(t_{u+1}^\theta - t_u^\theta)\right)\right\}$$

at each observation time t_j, $j = 1, \cdots, p$. After generating $U_i \sim U(0,1)$, if $F_{ij} \leqslant U_i < F_{i,j+1}$, we solved the following equation for t:

$$U_i = 1 - \exp\left\{-\nu^\theta\left(\sum_{i=0}^{j-1}\exp(\beta x(t_i))(t_{i+1}^\theta - t_i^\theta) + \exp(\beta x(t_j))(t^\theta - t_j^\theta)\right)\right\}. \qquad \text{(A.3)}$$

Then the solution of (A.3) will be the survival time, which is given by

$$T_i^0 = \left\{t_j^\theta - \exp(-\beta x(t_j))\left(\nu^{-\theta}\log(1-U_i) + \sum_{i=0}^{j-1}\exp(\beta x(t_i))(t_{i+1}^\theta - t_i^\theta)\right)\right\}^{\frac{1}{\theta}}. \qquad \text{(A.4)}$$

If $U_i > F_{i,p}$, then we set T_i^0 to be a big constant $M > t^*$. The follow up time $T_i = \min(T_i^0, V_i, t^*)$ and $D_i = I(T_i^0 \leqslant \min(V_i, t^*))$.

A.3 *The results with AR(1) correlation structure for time-varying exposures*

Table 5 presents the results for the AR(1) covariance structure using $\rho_{I_{AR}} = 0.938, 0.978,$ 0.996, which can be compared with the results in Table 1 for the CS covariance structure in the paper.

[Table 5 about here.]

B. Asymptotic distribution theory for $\hat{\boldsymbol{\beta}}_{rrc}$

B.1 *Approximate consistency of $\hat{\boldsymbol{\beta}}_{rrc}$*

We assume the following regularity conditions:

1. $\sup_{t\in[0,t^*]}\|\hat{\boldsymbol{\alpha}}_0(t) - \boldsymbol{\alpha}_0(t)\| \xrightarrow{p} \mathbf{0}$, $\sup_{t\in[0,t^*]}\|\hat{\boldsymbol{\alpha}}_1(t) - \boldsymbol{\alpha}_1(t)\| \xrightarrow{p} \mathbf{0}$, $\sup_{t\in[0,t^*]}\|\hat{\boldsymbol{\alpha}}_2(t) - \boldsymbol{\alpha}_2(t)\| \xrightarrow{p} \mathbf{0}$.
2. $s^{(0)}(\boldsymbol{\beta}, t)$, $\mathbf{s}^{(1)}(\boldsymbol{\beta}, t)$ and $\mathbf{s}^{(2)}(\boldsymbol{\beta}, t)$ are continuous functions of $\boldsymbol{\beta} \in \mathcal{B}$, uniformly in $t \in [0, t^*]$. $s^{(0)}(\boldsymbol{\beta}, t)$, $\mathbf{s}^{(1)}(\boldsymbol{\beta}, t)$ and $\mathbf{s}^{(2)}(\boldsymbol{\beta}, t)$ are bounded on $\mathcal{B} \times [0, t^*]$; $s^{(0)}(\boldsymbol{\beta}, t)$ is bounded away from zero on $\mathcal{B} \times [0, t^*]$.
3. Define

$$\mathbf{S}^{(2)}(\boldsymbol{\beta}, t) = {n_1}^{-1} \sum_{i=1}^{n_1} Y_m(i, t) \begin{pmatrix} \hat{\mathbf{x}}_i(t) \\ \mathbf{Z}_i(t) \end{pmatrix}^{\otimes 2} \exp\{\boldsymbol{\beta}_1^t \hat{\mathbf{x}}_i(t) + \boldsymbol{\beta}_2^t \mathbf{Z}_i(t)\},$$

then for $j = 0, 1, 2$, $\sup_{t\in[0,t^*],\beta\in\mathcal{B}}\|\mathbf{S}^{(j)}(\boldsymbol{\beta}, t) - \mathbf{s}^{(j)}(\boldsymbol{\beta}, t)\| \xrightarrow{p} \mathbf{0}$. For a vector v, we denote $v^{\otimes 0} = 1$, $v^{\otimes 1} = v$, $v^{\otimes 2} = vv'$.

Denote the left-hand side of equation (6) as $\mathbf{U}(\boldsymbol{\beta})$ and notice that $\mathbf{U}(\boldsymbol{\beta}) = \partial \mathbf{L}(\boldsymbol{\beta})/\partial \boldsymbol{\beta}$, where $\mathbf{L}(\boldsymbol{\beta})$ is the log-likelihood function with the expression:

$$n_1^{-1}\mathbf{L}(\boldsymbol{\beta}) = n_1^{-1} \sum_{i=1}^{n_1} \int_0^{t^*} [(\boldsymbol{\beta}_1^t \hat{\mathbf{x}}_i(t) + \boldsymbol{\beta}_2^t \mathbf{Z}_i(t)) - \log\{S^{(0)}(\boldsymbol{\beta}, \hat{\boldsymbol{\psi}}, t)\}] N_i(dt).$$

We can show that, under the regularity condition 1 - 3, $n_1^{-1}\mathbf{L}(\boldsymbol{\beta}) \xrightarrow{P} \mathbf{H}(\boldsymbol{\beta})$ with

$$\mathbf{H}(\boldsymbol{\beta}) = \int_0^{t^*} [\boldsymbol{\beta}^t \mathbf{s}^{(1)}(t) - \log\{s^{(0)}(\boldsymbol{\beta}, t)\} s^{(0)}(t)]\, dt$$

for each $\boldsymbol{\beta}$ in its parameter space $\mathcal{B}$, with $s^{(m)}(\boldsymbol{\beta}, t)$ and $s^{(m)}(t)$ defined as follows:

$$s^{(m)}(\boldsymbol{\beta}, t) = \mathrm{E}\left(Y_m(t) \begin{pmatrix} \tilde{\mathbf{x}}_i(\hat{\boldsymbol{\psi}}, t) \\ \mathbf{Z}_i(t) \end{pmatrix}^{\otimes m} \exp\{\boldsymbol{\beta}_1^t \tilde{\mathbf{x}}(t) + \boldsymbol{\beta}_2^t \mathbf{Z}(t)\} \right),$$

where $\tilde{\mathbf{x}}(t) = \boldsymbol{\alpha}_0(t) + \boldsymbol{\alpha}_1(t)\mathbf{X}(t) + \boldsymbol{\alpha}_2(t)\mathbf{Z}(t)$, and

$$s^{(m)}(t) = \lambda_0(t)\mathrm{E}\left[Y_m(t) \begin{pmatrix} \tilde{\mathbf{x}}_i(\hat{\boldsymbol{\psi}}, t) \\ \mathbf{Z}_i(t) \end{pmatrix}^{\otimes m} \mathrm{E}\left\{\exp(\boldsymbol{\beta}_{01}^t \mathbf{x}(t) + \boldsymbol{\beta}_{02}^t \mathbf{Z}(t)) | T \geqslant t, \mathbf{X}(t), \mathbf{Z}(t)\right\} \right],$$

where $m = 0, 1, 2$, $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{01}, \boldsymbol{\beta}_{02})$ is the true value of $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$.

Then, the first derivative, $\mathbf{h}(\boldsymbol{\beta}) \doteq \partial \mathbf{H}(\boldsymbol{\beta})/\partial \boldsymbol{\beta}$, is

$$\mathbf{h}(\boldsymbol{\beta}) = \int_0^{t^*} [\mathbf{s}^{(1)}(t) - \{\mathbf{s}^{(1)}(\boldsymbol{\beta}, t)/s^0(\boldsymbol{\beta}, t)\} s^{(0)}(t)]\, dt$$

and the second derivative, $-\mathbf{I}(\boldsymbol{\beta}) \doteq \partial^2 \mathbf{H}(\boldsymbol{\beta})/\partial \boldsymbol{\beta}^2$, is

$$-\mathbf{I}(\boldsymbol{\beta}) = -\int_0^{t^*} \left[\frac{\mathbf{s}^{(2)}(\boldsymbol{\beta}, t)}{\mathbf{s}^{(0)}(\boldsymbol{\beta}, t)} - \left\{ \frac{\mathbf{s}^{(1)}(\boldsymbol{\beta}, t)}{\mathbf{s}^{(0)}(\boldsymbol{\beta}, t)} \right\}^{\otimes 2} \right] s^{(0)}(t)\, dt.$$

We assume $\mathbf{I}(\boldsymbol{\beta})$ is positive definite, then the second derivative is negative definite. Set $\mathbf{h}(\boldsymbol{\beta}^*) = 0$, thus $\mathbf{H}(\boldsymbol{\beta})$ is a concave function with a unique maximum at $\boldsymbol{\beta} = \boldsymbol{\beta}^*$. Since $\hat{\boldsymbol{\beta}}_{rrc}$ maximizes the concave function $n_1^{-1}\mathbf{L}(\boldsymbol{\beta})$, by convex analysis (Andersen and Gill, 1982), we have $\hat{\boldsymbol{\beta}}_{rrc} \xrightarrow{P} \boldsymbol{\beta}^*$.

B.2 *Asymptotic normality of* $\hat{\boldsymbol{\beta}}_{rrc}$

Since the regression coefficients $\boldsymbol{\psi}(t) = (\boldsymbol{\alpha}_0(t), \boldsymbol{\alpha}_1(t), \boldsymbol{\alpha}_2(t))$ are estimated from the validation study, the variability of these estimates needs to be taken into account. We write the score equation (6) as $\mathbf{U}(\boldsymbol{\beta}, \boldsymbol{\psi})$ to indicate explicitly the dependence on $\boldsymbol{\psi}(t)$. Denote the true value of $\boldsymbol{\psi}(t)$ by $\boldsymbol{\psi}_0(t)$, which is now estimated by $\hat{\boldsymbol{\psi}}(t)$. Then, our estimating equation (6) is now

$\mathbf{U}(\hat{\boldsymbol{\beta}}_{rrc}, \hat{\boldsymbol{\psi}}) = 0$. Using Taylor's theorem, we can write

$$\mathbf{0} = \mathbf{U}(\hat{\boldsymbol{\beta}}_{rrc}, \hat{\boldsymbol{\psi}}) \approx \mathbf{U}(\boldsymbol{\beta}^*, \boldsymbol{\psi}_0) + \frac{\partial \mathbf{U}(\boldsymbol{\beta}^*, \boldsymbol{\psi}_0)}{\partial \boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}_{rrc} - \boldsymbol{\beta}^*) + \frac{\partial \mathbf{U}(\boldsymbol{\beta}^*, \boldsymbol{\psi}_0)}{\partial \boldsymbol{\psi}}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0).$$

Then,

$$n_1^{\frac{1}{2}}(\hat{\boldsymbol{\beta}}_{rrc} - \boldsymbol{\beta}^*) \approx \left[-n_1^{-1} \cdot \frac{\partial \mathbf{U}(\boldsymbol{\beta}^*, \boldsymbol{\psi}_0)}{\partial \boldsymbol{\beta}}\right]^{-1} \cdot n_1^{-\frac{1}{2}} \left[\mathbf{U}(\boldsymbol{\beta}^*, \boldsymbol{\psi}_0) + \frac{\partial \mathbf{U}(\boldsymbol{\beta}^*, \boldsymbol{\psi}_0)}{\partial \boldsymbol{\psi}}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0)\right].$$

Set

$$\hat{\mathbf{I}}(\boldsymbol{\beta}^*) = -n_1^{-1}\frac{\partial \mathbf{U}(\boldsymbol{\beta}^*, \boldsymbol{\psi}_0)}{\partial \boldsymbol{\beta}} = n_1^{-1}\sum_{i=1}^{n_1}\int_0^{t^*}\left[\frac{\mathbf{S}^{(2)}(\boldsymbol{\beta}^*, t)}{\mathbf{S}^{(0)}(\boldsymbol{\beta}^*, t)} - \left\{\frac{\mathbf{S}^{(1)}(\boldsymbol{\beta}^*, t)}{\mathbf{S}^{(0)}(\boldsymbol{\beta}^*, t)}\right\}^{\otimes 2}\right] N_i(dt),$$

then it can be easily verified that $\hat{\mathbf{I}}(\boldsymbol{\beta}^*) \xrightarrow{P} \mathbf{I}(\boldsymbol{\beta}^*)$ by following the proof in Anderson and Gill(1982). The matrix $\hat{\mathbf{I}}(\boldsymbol{\beta}^*)$ can be estimated by $\hat{\mathbf{I}}_{\boldsymbol{\beta}}$ in (12).

Also, it can be shown by following an argument similar to one used in the proof of theorem 2.1 in Lin and Wei (1989), that $n_1^{-\frac{1}{2}}\mathbf{U}(\boldsymbol{\beta}^*)$ is asymptotically equivalent to $n_1^{-\frac{1}{2}}\sum_{i=1}^{n_1}\mathbf{G}_i(\boldsymbol{\beta}^*)$, where

$$\mathbf{G}_i(\boldsymbol{\beta}^*) = \int_0^{t^*}\left\{\begin{pmatrix}\hat{\mathbf{x}}_i(t)\\ \mathbf{Z}_i(t)\end{pmatrix} - \frac{\mathbf{s}^{(1)}(\boldsymbol{\beta}^*, t)}{s^{(0)}(\boldsymbol{\beta}^*, t)}\right\} N_i(dt)$$
$$- \int_0^{t^*}\frac{Y_m(i,t)\exp(\boldsymbol{\beta}_1^*\hat{\mathbf{x}}_i(t) + \boldsymbol{\beta}_2^*\mathbf{Z}_i(t))}{s^{(0)}(\boldsymbol{\beta}^*, t)}\left\{\begin{pmatrix}\hat{\mathbf{x}}_i(t)\\ \mathbf{Z}_i(t)\end{pmatrix} - \frac{\mathbf{s}^{(1)}(\boldsymbol{\beta}^*, t)}{s^{(0)}(\boldsymbol{\beta}^*, t)}\right\}\tilde{F}(dt)$$

with $\tilde{F}(t) = \mathrm{E}(\sum_{i=1}^{n_1} N_i(t)/n_1)$.

So $n_1^{-\frac{1}{2}}\mathbf{U}(\boldsymbol{\beta}^*) \xrightarrow{D} N(\mathbf{0}, \mathbf{M}_1(\boldsymbol{\beta}^*))$ by the multivariate central limit theorem, with $\mathbf{M}_1(\boldsymbol{\beta}^*) = \mathrm{E}(\mathbf{G}_i(\boldsymbol{\beta}^*)^{\otimes 2})$, which can be estimated by $\hat{\mathbf{H}}_{\boldsymbol{\beta}}$ in (13).

To show the asymptotic normality of $\hat{\boldsymbol{\psi}}$, denote the left-hand side of (8) as $\mathbf{U}_{\boldsymbol{\psi}}(\boldsymbol{\psi})$. Then $\mathbf{U}_{\boldsymbol{\psi}}(\hat{\boldsymbol{\psi}}) = \mathbf{0}$. By the Taylor theorem, we have

$$\mathbf{0} = \mathbf{U}_{\boldsymbol{\psi}}(\hat{\boldsymbol{\psi}}) \approx \mathbf{U}_{\boldsymbol{\psi}}(\boldsymbol{\psi}_0) + \frac{\partial \mathbf{U}_{\boldsymbol{\psi}}}{\partial \boldsymbol{\psi}}(\boldsymbol{\psi}_0)(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0),$$

and it follows that

$$n_2^{\frac{1}{2}}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0) \approx -\left[\frac{1}{n_2}\frac{\partial \mathbf{U}_{\boldsymbol{\psi}}}{\partial \boldsymbol{\psi}}(\boldsymbol{\psi}_0)\right]^{-1} n_2^{-\frac{1}{2}}\mathbf{U}_{\boldsymbol{\psi}}(\boldsymbol{\psi}_0)$$

Hence, similar reasoning shows that $n_2^{\frac{1}{2}}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0) \xrightarrow{D} N(\mathbf{0}, \mathbf{M}_2(\boldsymbol{\psi}_0))$ and $\mathbf{M}_2(\boldsymbol{\psi}_0)$ can be estimated by $\hat{\mathbf{V}}_{\hat{\boldsymbol{\psi}}}$ in (10).

Therefore, $n_1^{\frac{1}{2}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$ is asymptotically normal with zero mean and covariance matrix $V(\boldsymbol{\beta}^*) = \hat{\mathbf{I}}(\boldsymbol{\beta}^*)^{-1}\tilde{\mathbf{M}}(\boldsymbol{\beta}^*)\hat{\mathbf{I}}(\boldsymbol{\beta}^*)^{-1}$, with $\tilde{\mathbf{M}}(\boldsymbol{\beta}^*) = \mathbf{M}_1(\boldsymbol{\beta}^*) + \frac{1}{n_1 n_2}\frac{\partial \mathbf{U}(\boldsymbol{\beta}^*, \boldsymbol{\psi}_0)}{\partial \boldsymbol{\psi}}\mathbf{M}_2(\boldsymbol{\psi}_0)(\frac{\partial \mathbf{U}(\boldsymbol{\beta}^*, \boldsymbol{\psi}_0)}{\partial \boldsymbol{\psi}})^T$. $V(\boldsymbol{\beta}^*)$ can be consistently estimated by $\hat{\mathbf{I}}_{\boldsymbol{\beta}}^{-1}\hat{\mathbf{H}}_{\boldsymbol{\beta},\boldsymbol{\psi}}\hat{\mathbf{I}}_{\boldsymbol{\beta}}^{-1}$ in (11).

20 November 2009

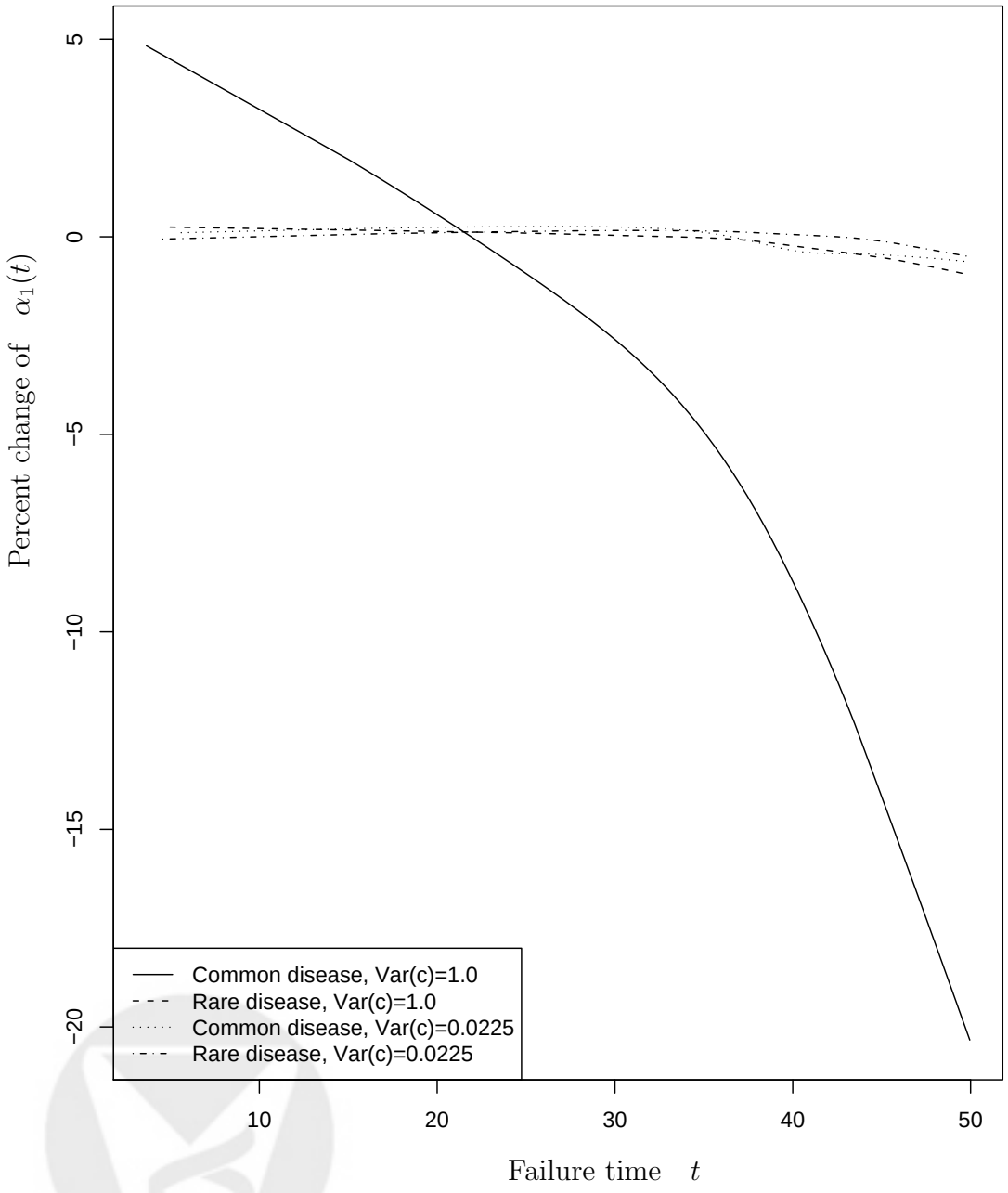

Figure 1. Plots were based on $\hat{\alpha}_1(t)$ from the conditional error model simulation with both $\mathrm{Var}(c) = 1.0$ and $\mathrm{Var}(c) = 0.0225$ scenario, $\rho = \mathrm{Corr}(c, C) = 0.3$.

Table 1

Results for simulation of cumulatively updated average exposure with a compound symmetry covariance structure, for different intra-class correlation $\rho_{I_{CS}}$.

$\rho_{I_{CS}}$	ρ	Estimated $\hat{\beta}(\hat{SE}[\hat{\beta}])$ Naive	RRC	Percent Bias(%) Naive	RRC	95% CI Coverage(%) Naive	RRC
		$n_1 = 50000$,	$n_2 = 150$,	Rare disease			
0.3	0.3	0.117(0.036)	0.502(0.179)	-76.6	0.4	0.0	95.6
	0.6	0.318(0.058)	0.500(0.098)	-36.3	0.1	12.8	94.1
	0.9	0.464(0.070)	0.499(0.077)	-7.3	-0.3	91.2	95.6
0.6	0.3	0.172(0.032)	0.509(0.118)	-65.6	1.7	0.0	95.0
	0.6	0.373(0.048)	0.498(0.069)	-25.4	-0.4	24.5	95.6
	0.9	0.474(0.054)	0.495(0.057)	-5.1	-0.9	92.6	94.9
0.9	0.3	0.212(0.030)	0.502(0.090)	-57.5	0.3	0.0	94.2
	0.6	0.405(0.041)	0.503(0.056)	-19.1	0.6	37.4	95.0
	0.9	0.486(0.045)	0.501(0.047)	-2.9	0.2	92.8	94.5
		$n_1 = 50000$,	$n_2 = 500$,	Rare disease			
0.3	0.3	0.119(0.036)	0.501(0.157)	-76.2	0.2	0.0	94.5
	0.6	0.313(0.058)	0.489(0.093)	-37.3	-2.1	9.0	95.2
	0.9	0.460(0.070)	0.494(0.076)	-8.0	-1.1	90.8	93.7
0.6	0.3	0.172(0.033)	0.499(0.101)	-65.6	-0.2	0.0	95.0
	0.6	0.374(0.048)	0.499(0.066)	-25.2	-0.2	24.6	95.2
	0.9	0.474(0.054)	0.495(0.057)	-5.2	-0.9	91.6	94.4
0.9	0.3	0.215(0.030)	0.506(0.077)	-57.0	1.2	0.0	94.9
	0.6	0.403(0.041)	0.498(0.052)	-19.5	-0.3	34.2	94.4
	0.9	0.486(0.045)	0.501(0.047)	-2.9	0.2	94.4	95.4
		$n_1 = 1000$,	$n_2 = 150$,	Common disease			
0.3	0.3	0.105(0.035)	0.492(0.193)	-79.1	-1.5	0.0	94.5
	0.6	0.293(0.058)	0.490(0.103)	-41.3	-2.1	4.7	94.4
	0.9	0.438(0.071)	0.490(0.078)	-12.5	-2.0	87.3	95.6
0.6	0.3	0.153(0.033)	0.509(0.135)	-69.4	1.8	0.0	94.1
	0.6	0.352(0.049)	0.503(0.076)	-29.6	0.7	14.3	94.8
	0.9	0.457(0.057)	0.498(0.061)	-8.5	-0.5	88.2	95.1
0.9	0.3	0.185(0.031)	0.502(0.107)	-62.9	0.3	0.0	94.3
	0.6	0.380(0.044)	0.504(0.063)	-24.1	0.8	22.0	94.1
	0.9	0.466(0.049)	0.501(0.052)	-6.8	0.1	89.8	94.6
		$n_1 = 1000$,	$n_2 = 500$,	Common disease			
0.3	0.3	0.103(0.035)	0.483(0.167)	-79.3	-3.3	0.0	94.7
	0.6	0.297(0.058)	0.497(0.096)	-40.7	-0.7	6.1	95.4
	0.9	0.440(0.071)	0.492(0.077)	-12.1	-1.7	86.9	94.7
0.6	0.3	0.152(0.033)	0.498(0.113)	-69.6	-0.4	0.0	93.8
	0.6	0.348(0.049)	0.497(0.071)	-30.5	-0.5	12.5	93.9
	0.9	0.460(0.057)	0.501(0.060)	-8.0	0.1	89.9	95.0
0.9	0.3	0.187(0.031)	0.504(0.089)	-62.5	0.8	0.0	96.2
	0.6	0.378(0.044)	0.503(0.059)	-24.3	0.5	21.8	93.9
	0.9	0.464(0.049)	0.498(0.051)	-7.3	-0.4	88.3	95.5

True $\beta = 0.5$, the study duration $t^* = 50$, the number of simulation replications $B = 1000$.
In the rare disease situation, the cumulative incidence is about 1% with $n_1 = 50000$.
In the common disease situation, the cumulative incidence is about 50% with $n_1 = 1000$.

Table 2
Basic characteristics of the study population.

Variable	Exposure type	Validation study						Main study	
		HPFS ($n_{21} = 127$, Age:(39 - 75))			EATS ($n_{22} = 446$, Age:(21 - 76))			HPFS($n_1 = 47760$)	
		Mean(s.d.)	Range	Correlation	Mean(s.d.)	Range	Correlation	Mean(s.d.)	Range
Total calcium[a]	DR	0.43(0.12)	(0.24, 0.87)		0.46(0.16)	(0.18, 1.02)		-	-
(1838 mg/day)	FFQ	0.47(0.17)	(0.25, 1.17)	0.51	0.53(0.22)	(0.14, 1.54)	0.61	0.50(0.21)	(0.07,3.96)
Total vitamin E[b]	DR	1.32(0.42)	(0.85, 2.72)		1.36(0.56)	(0.57, 3.03)		-	-
(mg/day)	FFQ	1.20(0.49)	(0.71, 2.97)	0.89	1.33(0.53)	(0.33, 2.87)	0.79	1.42(0.54)	(0.48,3.02)

[a]Total calcium intake is energy adjusted and the unit is 1838 mg/day.
[b]Total vitamin E intake is energy adjusted and $\log_{10}$ transformed.

Table 3

Estimated coefficients and standard errors for HPFS of the relationship between total calcium intake and the fatal prostate cancer incidence.

		Total Calcium intake[a]			Total Vitamin E intake[b]		
Method	Validation study	Estimate(S.E.)	RR(95% C.I.)	p-value	Estimate(S.E.)	RR(95% C.I.)	p-value
Uncorrected	N/A	0.587(0.239)	1.80[1.13,2.87]	0.014	-0.274(0.103)	0.76[0.62,0.93]	0.008
RRC,	HPFS	0.623(0.653)	1.87[0.52,6.71]	0.340	-0.342(0.160)	0.71[0.52,0.97]	0.033
Vit-E error-free	HPFS + EATS	1.359(0.500)	3.89[1.46,10.37]	0.007	-0.374(0.120)	0.69[0.54,0.87]	0.002
RRC, Cal,Vit-E	HPFS	0.150(0.766)	1.16[0.26,5.22]	0.845	-0.203(0.447)	0.82[0.34,1.96]	0.650
both error-prone	HPFS + EATS	0.734(0.371)	2.08[1.01,4.31]	0.048	-0.370(0.201)	0.69[0.47,1.02]	0.065

[a] Total calcium intake is energy adjusted and the unit is 1838 mg/day.

[b] Total vitamin E intake is energy adjusted and $\log_{10}$ transformed.

Table 4

Results for simulation of time-invariant exposure with a conditional normal error model, for different correlation ρ between c and C.

	Estimated $\hat{\beta}(\hat{SE}[\hat{\beta}])$			Percent Bias(%)			95% CI Coverage(%)		
ρ	Naive	ORC	RRC	Naive	ORC	RRC	Naive	ORC	RRC
	Parameters :	$\mathrm{E}(c) = 0.45$,	$\mathrm{Var}(c) = 0.0225$,	$\mathrm{E}(C) = 0.5$,	$\mathrm{Var}(C) = 0.04$				
	$n_1 = 50000$,	$n_2 = 150$,	Rare disease						
0.3	0.114(0.223)	0.590(1.361)	0.559(1.313)	-77.2	18.0	11.7	59.9	96.9	98.1
0.6	0.226(0.223)	0.509(0.509)	0.512(0.514)	-54.7	1.7	2.4	76.2	95.5	95.4
0.9	0.339(0.223)	0.504(0.333)	0.505(0.333)	-32.2	0.7	1.0	88.7	95.2	94.7
	$n_1 = 50000$,	$n_2 = 500$,	Rare disease						
0.3	0.114(0.223)	0.524(1.029)	0.539(1.052)	-77.2	4.9	7.9	58.1	95.9	96.6
0.6	0.224(0.223)	0.500(0.500)	0.501(0.502)	-55.2	0.0	0.3	75.6	96.4	96.7
0.9	0.335(0.223)	0.497(0.331)	0.497(0.331)	-33.0	-0.6	-0.6	87.7	95.5	95.3
	$n_1 = 1000$,	$n_2 = 150$,	Common disease						
0.3	0.109(0.224)	0.503(1.112)	0.512(1.275)	-78.1	0.6	2.4	58.7	97.6	98.6
0.6	0.230(0.225)	0.520(0.513)	0.523(0.527)	-54.0	4.0	4.6	78.2	95.2	95.9
0.9	0.341(0.225)	0.505(0.334)	0.507(0.336)	-31.7	1.0	1.3	88.9	94.7	95.4
	$n_1 = 1000$,	$n_2 = 500$,	Common disease						
0.3	0.115(0.225)	0.515(1.034)	0.530(1.086)	-77.1	3.1	6.0	59.4	95.7	96.7
0.6	0.227(0.224)	0.508(0.503)	0.509(0.506)	-54.5	1.6	1.7	77.0	95.9	96.0
0.9	0.334(0.225)	0.494(0.333)	0.494(0.334)	-33.3	-1.2	-1.1	88.3	94.7	95.0
	Parameters :	$\mathrm{E}(c) = 0.45$,	$\mathrm{Var}(c) = 1.0$,	$\mathrm{E}(C) = 0.5$,	$\mathrm{Var}(C) = 2.0$				
	$n_1 = 50000$,	$n_2 = 150$,	Rare disease						
0.3	0.107(0.032)	0.557(0.316)	0.566(0.327)	-78.6	11.4	13.2	0.0	94.8	93.8
0.6	0.211(0.032)	0.505(0.095)	0.507(0.103)	-57.7	0.9	1.5	0.0	96.0	95.5
0.9	0.319(0.032)	0.502(0.054)	0.503(0.056)	-36.3	0.4	0.7	0.0	94.5	94.9
	$n_1 = 50000$,	$n_2 = 500$,	Rare disease						
0.3	0.105(0.032)	0.508(0.172)	0.516(0.185)	-79.0	1.6	3.2	0.0	95.8	95.8
0.6	0.212(0.032)	0.501(0.081)	0.503(0.084)	-57.6	0.2	0.6	0.0	95.8	95.8
0.9	0.318(0.032)	0.500(0.051)	0.500(0.052)	-36.5	-0.1	0.0	0.0	95.9	95.6
	$n_1 = 1000$,	$n_2 = 150$,	Common disease						
0.3	0.095(0.032)	0.474(0.220)	0.494(0.276)	-81.1	-5.3	-1.1	0.0	89.1	89.1
0.6	0.198(0.032)	0.473(0.095)	0.506(0.116)	-60.4	-5.3	1.3	0.0	91.6	93.8
0.9	0.311(0.033)	0.490(0.056)	0.500(0.061)	-37.7	-2.1	0.0	0.0	94.6	95.6
	$n_1 = 1000$,	$n_2 = 500$,	Common disease						
0.3	0.095(0.032)	0.457(0.170)	0.506(0.207)	-81.0	-8.6	1.3	0.0	93.6	95.1
0.6	0.198(0.032)	0.468(0.082)	0.500(0.093)	-60.5	-6.4	0.0	0.0	91.3	94.7
0.9	0.310(0.033)	0.487(0.053)	0.497(0.056)	-38.0	-2.6	-0.5	0.0	94.3	95.6

True $\beta = 0.5$, the study duration $t^* = 50$, the number of simulation replications $B = 1000$.
In the rare disease situation, the cumulative incidence is about 1% with $n_1 = 50000$.
In the common disease situation, the cumulative incidence is about 50% with $n_1 = 1000$.

Table 5

Results for simulation of cumulatively updated average exposure with an AR(1) covariance structure, for different intra-class correlation $\rho_{I_{AR}}$.

$\rho_{I_{AR}}$	ρ	Estimated $\hat{\beta}(\hat{SE}[\hat{\beta}])$ Naive	RRC	Percent Bias(%) Naive	RRC	95% CI Coverage(%) Naive	RRC
		$n_1 = 50000$,	$n_2 = 150$,	Rare disease			
0.938	0.3	0.202(0.031)	0.500(0.095)	-59.7	0.0	0.0	95.2
	0.6	0.396(0.043)	0.500(0.059)	-20.9	0.0	30.6	94.6
	0.9	0.482(0.047)	0.499(0.049)	-3.6	-0.3	92.7	94.8
0.978	0.3	0.216(0.030)	0.506(0.089)	-56.7	1.2	0.0	94.2
	0.6	0.405(0.040)	0.498(0.055)	-19.0	-0.5	33.7	95.7
	0.9	0.482(0.044)	0.497(0.046)	-3.6	-0.7	94.3	96.1
0.996	0.3	0.223(0.029)	0.503(0.085)	-55.4	0.7	0.0	94.4
	0.6	0.411(0.039)	0.504(0.053)	-17.8	0.7	39.5	94.6
	0.9	0.486(0.043)	0.501(0.045)	-2.7	0.1	93.2	94.5
		$n_1 = 50000$,	$n_2 = 500$,	Rare disease			
0.938	0.3	0.204(0.031)	0.501(0.081)	-59.2	0.3	0.0	94.5
	0.6	0.392(0.043)	0.492(0.055)	-21.7	-1.6	25.7	94.6
	0.9	0.481(0.047)	0.497(0.049)	-3.8	-0.5	92.1	93.5
0.978	0.3	0.216(0.030)	0.498(0.074)	-56.9	-0.5	0.0	94.5
	0.6	0.405(0.040)	0.499(0.051)	-18.9	-0.1	35.4	94.6
	0.9	0.481(0.044)	0.496(0.046)	-3.8	-0.8	92.4	93.7
0.996	0.3	0.225(0.029)	0.506(0.072)	-55.0	1.2	0.0	95.3
	0.6	0.409(0.039)	0.498(0.049)	-18.3	-0.3	37.1	94.2
	0.9	0.486(0.043)	0.501(0.044)	-2.7	0.1	94.2	95.1
		$n_1 = 1000$,	$n_2 = 150$,	Common disease			
0.938	0.3	0.178(0.031)	0.496(0.109)	-64.5	-0.9	0.0	93.7
	0.6	0.368(0.045)	0.493(0.065)	-26.3	-1.4	17.4	94.8
	0.9	0.460(0.050)	0.495(0.054)	-8.0	-1.0	87.9	96.0
0.978	0.3	0.190(0.031)	0.507(0.105)	-62.0	1.5	0.0	93.5
	0.6	0.382(0.044)	0.502(0.062)	-23.6	0.5	22.1	94.1
	0.9	0.464(0.048)	0.499(0.051)	-7.1	-0.3	88.9	95.4
0.996	0.3	0.194(0.030)	0.501(0.101)	-61.3	0.1	0.0	93.9
	0.6	0.385(0.043)	0.503(0.061)	-22.9	0.7	23.1	94.4
	0.9	0.467(0.047)	0.501(0.050)	-6.6	0.1	89.4	94.3
		$n_1 = 1000$,	$n_2 = 500$,	Common disease			
0.938	0.3	0.176(0.031)	0.489(0.092)	-64.7	-2.2	0.0	94.6
	0.6	0.370(0.045)	0.498(0.060)	-25.9	-0.5	17.4	95.6
	0.9	0.461(0.050)	0.496(0.053)	-7.8	-0.8	88.1	94.4
0.978	0.3	0.189(0.031)	0.497(0.086)	-62.1	-0.6	0.0	94.1
	0.6	0.378(0.044)	0.498(0.057)	-24.4	-0.4	19.9	93.7
	0.9	0.467(0.048)	0.501(0.051)	-6.7	0.2	89.1	94.1
0.996	0.3	0.195(0.030)	0.503(0.084)	-60.9	0.6	0.0	96.3
	0.6	0.384(0.043)	0.502(0.056)	-23.3	0.3	23.4	93.3
	0.9	0.465(0.047)	0.498(0.049)	-7.0	-0.3	88.2	95.3

True $\beta = 0.5$, the study duration $t^* = 50$, the number of simulation replications $B = 1000$.
In the rare disease situation, the cumulative incidence is about 1% with $n_1 = 50000$.
In the common disease situation, the cumulative incidence is about 50% with $n_1 = 1000$.
$\rho_{I_{AR}} = 0.938, 0.978, 0.996$ are respectively in an equal footing with $\rho_{I_{CS}} = 0.3, 0.6, 0.9$ according to the equation (15).