Harvard University
Harvard University Biostatistics Working Paper Series

Year 2015 *Paper* 198

Survival analysis with functions of mis-measured covariate histories: the case of chronic air pollution exposure in relation to mortality in the Nurses' Health Study

Xiaomei Liao* Molin Wang† Jaime E. Hart‡
Francine Laden** Donna Spiegelman††

*Harvard University, stxia@channing.harvard.edu
†Harvard University
‡Harvard University
**Harvard University
††Harvard University

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.

http://biostats.bepress.com/harvardbiostat/paper198

Copyright ©2015 by the authors.

Survival analysis with functions of mis-measured covariate histories: the case of chronic air pollution exposure in relation to mortality in the Nurses' Health Study

Xiaomei Liao, Molin Wang, Jaime E. Hart, Francine Laden, and Donna Spiegelman

Abstract

Environmental epidemiologists are often interested in estimating the effect of functions of time-varying exposure histories, such as the 12-month moving average, in relation to chronic disease incidence or mortality. The individual exposure measurements that comprise such an exposure history are usually mis-measured, at least moderately, and, often, more substantially. To obtain unbiased estimates of Cox model hazard ratios for these complex mis-measured exposure functions, an extended risk set regression calibration (RRC) method for Cox models is developed and applied to a study of long-term exposure to the fine particulate matter ($PM_{2.5}$) component of air pollution in relation to all-cause mortality in the Nurses' Health Study. Simulation studies under several realistic assumptions about the measurement error model and about the correlation structure of the repeated exposure measurements were conducted to assess the finite sample properties of this new method, and found that the method has good performance in terms of finite sample bias reduction and nominal confidence interval coverage. User-friendly software has been developed and is available to the general public on the senior author's website.

Survival analysis with functions of mis-measured covariate histories: the case of chronic air pollution exposure in relation to mortality in the Nurses' Health Study

Xiaomei Liao*[1,2,4], Molin Wang[1,2,4], Jaime E. Hart[3,4], Francine Laden[2,3,4], and Donna Spiegelman[1,2,4,5,6]

[1]Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, MA, 02115, USA.
[2]Department of Epidemiology, Harvard T. H. Chan School of Public Health, Boston, MA, 02115, USA.
[3]Department of Environmental Health, Harvard T. H. Chan School of Public Health, Boston, MA, 02215, USA.
[4]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, 02115, USA.
[5]Department of Nutrition, Harvard T. H. Chan School of Public Health, Boston, MA, 02115, USA.
[6]Department of Global Health and Population, Harvard T. H. Chan School of Public Health, Boston, MA, 02115, USA.

July 16, 2015

Abstract

Environmental epidemiologists are often interested in estimating the effect of functions of time-varying exposure histories, such as the 12-month moving average, in relation to chronic disease incidence or mortality. The individual exposure measurements that comprise such an exposure history are usually mis-measured, at least moderately, and, often, more substantially. To obtain unbiased estimates of Cox model hazard ratios for these complex mis-measured exposure functions, an

*Email: stxia@channing.harvard.edu. This research was supported by NIH/NIEHS grant R01ES09411.

Hosted by The Berkeley Electronic Press

extended risk set regression calibration (RRC) method for Cox models is developed and applied to a study of long-term exposure to the fine particulate matter ($PM_{2.5}$) component of air pollution in relation to all-cause mortality in the Nurses' Health Study. Simulation studies under several realistic assumptions about the measurement error model and about the correlation structure of the repeated exposure measurements were conducted to assess the finite sample properties of this new method, and found that the method has good performance in terms of finite sample bias reduction and nominal confidence interval coverage. User-friendly software has been developed and is available to the general public on the senior author's website.

Keywords: Functions of time-varying exposure histories; Measurement error; Cox proportional hazards model; Risk set regression calibration.

1 INTRODUCTION

Air pollution (AP) is the 13th leading cause of mortality worldwide (Brook et al., 2004). The acute effects of AP on mortality have been well documented as far back as 1930 (Pope III, 2000) and evidence indicates that the relationship is both valid and causal (Dab et al., 2001). Since the late 1980s, numerous studies have found current particulate air pollutant concentrations to be responsible for excess mortality, with associations with daily deaths having been reported across the U.S., Europe, Latin America and Asia (Atkinson et al., 2015, 2014; U.S. EPA., 2004). In addition, a large body of evidence has formed linking long-term, that is chronic, AP exposure with an increased risk of cardiovascular (CVD) morbidity and mortality (Hoek et al., 2013; Pope III and Dockery, 2006). In a recent review of the evidence from 11 long-term AP cohort studies, the pooled effect estimate for each 10 $\mu g/m^3$ increase in $PM_{2.5}$ (particulate matter contained in AP of size fractions up to 2.5 micrometers (μm)) was 6.2% (95%CI: 4.1% - 8.4%) for all-cause mortality and 10.6% (95%CI: 5.4% -16.0%) for cardiovascular mortality (Hoek et al., 2013). To put these risks into perspective, in 1988 and 2013, average $PM_{2.5}$ levels in the U.S. were 35 $\mu g/m^3$ and under 10 $\mu g/m^3$ respectively, while currently in Beijing, China and Delhi, India, they are approximately 100 $\mu g/m^3$ and 150 $\mu g/m^3$ respectively.

In chronic AP epidemiology, the 12-month moving average has been commonly used

http://biostats.bepress.com/harvardbiostat/paper198

(Schwartz et al., 2008). Thus, in the study motivating this methodologic research, the AP exposure metrics of interest are the 12-month moving averages of $PM_{2.5}$, PM_{10} and $PM_{2.5-10}$, particulate matter contained in AP of size fractions up to 2.5 μm, up to 10 μm, and between 2.5 and 10 μm (Puett et al., 2008; Hart et al., 2015). The individual exposure data that are used to construct the 12-month $PM_{2.5}$ moving average exposure were collected monthly from the nearest Environmental Protection Agency (EPA) Air Quality System (AQS) monitor, and then estimated from well-established spatio-temporal models that smooth the EPA monitor measurements over time and space, taking other factors into account, such as meteorological conditions, as well (Paciorek et al., 2009; Yanosky et al., 2014). However, neither the measurements from EPA monitors nor the spatio-temporal smoothed exposure estimates accurately reflect personal exposure to particulate matter for each individual included in the cohort at each point in time during which they are under observation. A recent nine city validation study showed that these surrogate individual exposure measurements at a given point in time have moderate to substantial measurement error, with correlations between these surrogates and direct measures of personal exposure ranging from 0.29 to 0.60 (Kioumourtzoglou et al., 2014). It should be noted that the methodology developed here is fully generalizable to any pre-specified functions of the exposure histories.

To validly and efficiently assess the effect of environmental exposures on health outcomes, exposure measurement error needs to be taken into account or exposure effects will likely be under- and poorly estimated (Carroll et al., 2006). An additional challenge in the analysis of AP cohort studies, where the exposure of interest is a function of the mis-measured exposure history, concerns the mis-alignment of the validation study data with the data needed to correct the effect estimates to be obtained in the main study for measurement error, in the following sense: validation studies contain, at most, a small sample of individual point exposure measurements at several points in time, while, as

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

discussed above, the exposure of interest in the main study, is a pre-specified function of the exposure history. In the data motivating this research, each participant has a series of surrogate 12-month moving average exposure functions measured over some interval in their life time, each one of which imperfectly measures its unobserved perfectly measured counterpart. Validating a sample of these 12-month moving averages over the etiologically relevant period or even a reasonably sized sub-interval is not possible. Instead of validating the function of the surrogate exposure histories, exposure validation studies in AP research and, more generally, in environmental health research, are comprised of a sample of point exposures typically observed with few or no repeated measures within each subject. The methods developed below are designed to accommodate this challenging common setting.

The regression calibration (RC) method for survival data analysis was first proposed by Prentice (1982). Further research extended these initial results to a somewhat broader range of models and measurement error structures and to different study designs that permit consistent estimation and inference (Rosner et al., 1989, 1990; Spiegelman et al., 1997; Wang et al., 1997). The original regression calibration method for survival data analysis assumed that the measurement error model is time-invariant, as is approximately true when the disease is rare. Otherwise, if the rare disease assumption does not hold, a risk set regression calibration (RRC) method which allows the measurement error model to evolve over time and refits at each risk set can be used (Xie et al., 2001). Liao et al. (2011) further expanded these methods, and addressed measurement error correction for a limited class of time-varying covariates: time-varying point exposures, and functions of time-varying point exposures when, unrealistically, the validation study has the same frequency of measurement of time-varying point exposures as in the main study or by assuming, generally unrealistically, that the measurement error model for the point exposures in the validation study is fully transportable to the main study for the exposure metric of

COBRA
BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/harvardbiostat/paper198

interest, a time-varying function of the exposure histories of varying duration. Without making unrealistic assumptions as above, there is no method available to validly estimate hazard ratios for functions of time-varying mis-measured point exposures with limited validation data as in this case study.

This paper develops an approximately consistent method for estimating the hazard ratio for functions of time-varying mis-measured exposure histories, augmented by point exposure validation data. It is organized as follows. In Section 2, we introduce the motivating example. Then, we develop the new methodology in Section 3. We conduct a simulation study in Section 4. In Section 5, we apply the proposed method to the analysis of $PM_{2.5}$ in relation to all-cause mortality in the Nurses' Health Study (NHS), described in Section 2, augmented by the nine city validation study. In Section 6, we conclude with a discussion of the limitations and strengths of the proposed method, and identify directions for further research. User-friendly software accompanies the new method, and can be found on the senior author's website, `https://www.hsph.harvard.edu/donna-spiegelman/software/rrc-macro/`.

2 APPLICATION: NURSES' HEALTH STUDY AIR POLLUTION COHORT AND NINE CITY VALIDATION STUDY

2.1 Nurses' Health Study Air Pollution Cohort

This study of the chronic health effects of AP is embedded within the NHS, a long-term prospective cohort study initiated in 1976 when 121,701 married female US registered nurses, 30 to 55 years old, completed a mailed questionnaire (Colditz and Hankinson, 2005). While participants were initially from eleven states at the study's inception, the

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics

Hosted by The Berkeley Electronic Press

participants now reside in all 50 U.S. states. Follow-up questionnaires have been mailed every two years to obtain updated information on risk factors and on the occurrence of major illnesses, with response rates above 90% for each mailing, making it possible to regularly update residential addresses. This study includes the 117,408 participants who were alive in 1988, when estimates of particulate matter exposure were first available, and who had at least one address within the continental U.S. between 1988 and 2006 for which longitude and latitude could be obtained. Non-accidental deaths occuring between June 1988 through May 2006 were identified through state vital statistics records, the National Death Index, by report from the families and by the postal system (Colditz et al., 1986). A nationwide expansion of a geographic information system (GIS)-based spatio-temporal model for $PM_{2.5}$ was used to compute smoothed monthly $PM_{2.5}$ exposure at each participants' residential address (Paciorek et al., 2009; Yanosky et al., 2014). Based on the monthly measurements from June 1999 - May 2006, the 12-month moving average exposures were then calculated.

2.2 Nine City Validation Study

Ideally, $PM_{2.5}$ exposure would be measured by personal monitors continuously over the entire follow-up period of each participant. If this were possible, the only exposure measurement error would be technical error associated with the monitoring devices, which is small relative to the error between personal exposure measurements and the surrogates described above (Chakrabarti et al., 2004; Lee et al., 2006). Unfortunately, this method of exposure assessment is impossible: it would be prohibitively expensive and unacceptably intrusive to expect hundreds of thousands of free-living human subjects to agree to wear personal monitors over their entire adult life, while awake and while sleeping. In the absence of these 'gold standard' point exposure measurements over the participants' entire exposure history, validation data are necessary for valid estimation and inference of

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/harvardbiostat/paper198

relative risks. The available external validation study includes measurements of personal and ambient $PM_{2.5}$ between 1998-2002 from nine cities throughout the United States (Kioumourtzoglou et al., 2014) : Atlanta, GA; Baltimore, MD; Boston, MA; Los Angeles, CA; Research Triangle Park (RTP), NC; Elizabeth, NJ; Houston, TX; Seattle, WA; Steubenville, OH. The number of subjects in these studies ranged between 15 and 201, with sampling session durations between 2 to 12 days. The exposure of primary interest, *personal* exposure to $PM_{2.5}$ of *ambient* origin, corresponds to $PM_{2.5}$ from *outdoor* sources of AP only (Sheppard et al., 2005). This quantity cannot be directly assessed, but is estimated by adjusting the available measurements, which measure AP from both outdoor and indoor sources together, with a measurement of the ratio of indoor to outdoor SO_4^{2-} (Sarnat et al., 2002; Wilson et al., 2000). The latter data were available in 5 of the 9 cities included in the validation study.

In these validation data, two types of surrogate monthly $PM_{2.5}$ exposure measurements were available for each subject, just as in the NHS AP cohort — the monthly ambient $PM_{2.5}$ concentration from the nearest EPA AQS monitor and the smoothed monthly outdoor $PM_{2.5}$ concentration, using the nationally expanded spatio-temporal model (Paciorek et al., 2009; Yanosky et al., 2014) linked to participants' geocoded residential address. Both of these surrogate exposures measured personal exposure to $PM_{2.5}$ of ambient origin, the true point exposure of interest here, with a substantial amount of error; the correlations between these surrogates and the true point exposure ranged from 0.29 to 0.60. The calibration factors, estimated as the slopes from the regression of personal $PM_{2.5}$ of ambient origin on the surrogate exposures, were 0.31 for the nearest EPA monitor exposure and 0.54 for the spatio-temporal model predicted exposure (Kioumourtzoglou et al., 2014), suggesting that risk estimates will be attenuated if either of the surrogate exposures is used in the analysis, since only when the calibration factor is 1, there is no bias in the effect estimate (Carroll et al., 2006).

CODRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

After excluding younger subjects due to their age discrepancy with the NHS AP cohort, 40% of the validation study subjects had more than one pair of measurements of personal and surrogate exposures. The median and interquartile range (IQR) of the duration of time spanned by these repeated measurements were 6.0 and 5.0 months, respectively. Thus, even with the most detailed exposure validation data available for use with U.S. populations, it would not be possible to validate any function of the exposure history of interest, such as the 12-month moving average. The new method is developed to permit valid analysis of these data, making it possible to obtain accurate estimates of the hazard ratio using the methods developed below.

3 FUNCTIONS OF MIS-MEASURED TIME VARYING EXPOSURE HISTORIES

3.1 The Estimator $\hat{\boldsymbol{\beta}}_{corr}$

The model of interest here is the Cox model (Cox, 1972) for censored survival data. We consider the following hazard rate function for the survival time, T, of an individual:

$$\lambda(t; \{\mathbf{c}(v), v \leq t\}, \{\mathbf{W}(v), v \leq t\}) = \lambda(t; \mathbf{x}(t), \mathbf{Z}(t)) = \lambda_0(t) \exp(\boldsymbol{\beta}_x' \mathbf{x}(t) + \boldsymbol{\beta}_z' \mathbf{Z}(t)), \quad (1)$$

where $\lambda_0(t)$ is the baseline hazard function and $\boldsymbol{\beta} = (\boldsymbol{\beta}_x, \boldsymbol{\beta}_z)$ is a $(p+q)$-vector of regression coefficients. The p-dimensional time-varying exposure vector $\mathbf{x}(t)$ is a function of time-varying point exposure histories $\{\mathbf{c}(v), v \leq t\}$, which are subject to error, and the q-dimensional time-varying error-free covariate vector $\mathbf{Z}(t)$ is a function of the covariate histories $\{\mathbf{W}(v), v \leq t\}$. For example, $\mathbf{Z}(t)$ may be set to be the most recent measurement in the histories, i.e., $\mathbf{Z}(t) = \mathbf{W}(t)$, but often other functions of the histories may be needed to fully adjust for confounding by $\mathbf{W}(t)$.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/harvardbiostat/paper198

In AP research, the moving average exposure of $a-b$ duration $\mathbf{x}(t)$ is a major focus, and at time t_k of a discrete time scale $(t_0, t_1, \cdots, t_k, \cdots, t_m)$ is given by

$$\mathbf{x}(t_k) = \sum_{j=k-b+1}^{k-a+1} \frac{\mathbf{c}(t_j) \times I(t_j)}{\sum_{j=k-b+1}^{k-a+1} I(t_j)}, \tag{2}$$

where $I(t_j)$ is an indicator function for whether the exposure was measured at time t_j, and a, b are positive integers with $b \geq a$. For the 12-month moving average exposure as motivated by the data at hand, $a = 1$, $b = 12$ and the time scale is in months. Instead of observing the time-varying point exposure histories $\{\mathbf{c}(t), t_0 \leq t \leq t_m\}$, the NHS measures the surrogate exposure histories $\{\mathbf{C}(t), t_0 \leq t \leq t_m\}$, where, in particular, the surrogate $a-b$ moving average exposure is $\mathbf{X}(t_k) = \sum_{j=k-b+1}^{k-a+1} \frac{\mathbf{C}(t_j) \times I(t_j)}{\sum_{j=k-b+1}^{k-a+1} I(t_j)}$. As discussed more heuristically in the previous section, it is not possible to validate $\mathbf{X}(t)$ because it is a function of the individual histories of exposures and few exposures have been repeatedly validated in the same study participants. Hence, we propose a method for obtaining an estimate of $\mathbf{x}(t)$ from $\{\mathbf{C}(t), t_0 \leq t \leq t_m\}$ in the main study using the limited point exposure data available in the validation study. The corrected hazard ratio estimator follows by substituting $\hat{\mathbf{x}}(t)$ for $\mathbf{x}(t)$ in (1).

The method is developed for the main/external validation study design. The main study consists of data $\{\{\mathbf{C}_i(t), t_0 \leq t \leq t_m\}, \{\mathbf{W}_i(t), t_0 \leq t \leq t_m\}, T_i, D_i\}$, $i = 1, \cdots, n_1$, and $\{\mathbf{C}_i(t), \mathbf{W}_i(t)\}$ is measured at a discrete grid of time points $(t_0, t_1, \cdots, t_m)$. Here T_i is the follow-up time, which is defined as the minimum of the potential failure time T_i^0, the potential censoring time V_i, and the administrative end of follow-up, t^*, i.e., $T_i = min(T_i^0, V_i, t^*)$. The validation study consists of data $\{\mathbf{c}_i(t), \mathbf{C}_i(t), \mathbf{W}_i(t)\}$, $i = n_1 + 1, \cdots, n_1 + n_2$, while $\{\mathbf{c}_i(t), \mathbf{C}_i(t), \mathbf{W}_i(t)\}$ is measured less frequently, typically, as in our case study, no more than at two time points. If case status, D_i, is all-cause mortality as in the data motivating this research, all validation study participants are alive and censored, i.e. $D_i = 0$ for all the validation subjects, and T_i is known as well,

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

$i = n_1 + 1, \cdots, n_1 + n_2$.

In the main study/external validation study design, transportability needs to be assumed, that is, the parameters of measurement error model that are later used for bias correction can reasonably be assumed to be the same as those which produced the mismeasured data in the main study (Carroll et al., 2006). In addition, we make the standard conditional independence assumption of measurement error modeling methods,

$$\lambda(t; \{\mathbf{c}(v), v \le t\}, \{\mathbf{C}(v), v \le t\}, \{\mathbf{W}(v), v \le t\}) = \lambda(t; \{\mathbf{c}(v), v \le t\}, \{\mathbf{W}(v), v \le t\}), \quad (3)$$

implying that the surrogate exposure histories $\{\mathbf{C}(v), v \le t\}$ have no association with the outcome given the true exposure histories $\{\mathbf{c}(v), v \le t\}$ (Prentice, 1982). In addition, following Prentice (1982), we assume the censorship satifies $\lambda(t, \{\mathbf{C}(v), v \le t\}, \{\mathbf{W}(v), v \le t\}, \text{no censorship in } [0, t)) = \lambda(t, \{\mathbf{C}(v), v \le t\}, \{\mathbf{W}(v), v \le t\})$.

Then, it follows that

$$\begin{aligned} & \lambda(t; \{\mathbf{C}(v), v \le t\}, \{\mathbf{W}(v), v \le t\}) \\ = \; & \mathrm{E}_{\mathbf{c}} \left[\lambda(t; \{\mathbf{c}(v), v \le t\}, \{\mathbf{C}(v), v \le t\}, \{\mathbf{W}(v), v \le t\}) | T \ge t, \{\mathbf{C}(v), v \le t\}, \{\mathbf{W}(v), v \le t\} \right] \\ = \; & \mathrm{E}_{\mathbf{c}} \left[\lambda(t; \{\mathbf{c}(v), v \le t\}, \{\mathbf{W}(v), v \le t\}) | T \ge t, \{\mathbf{C}(v), v \le t\}, \{\mathbf{W}(v), v \le t\} \right], \end{aligned}$$

or equivalently,

$$\begin{aligned} & \lambda(t; \{\mathbf{C}(v), v \le t\}, \{\mathbf{W}(v), v \le t\}) \\ = \; & \mathrm{E}_{\mathbf{x}(t)} \left[\lambda(t; \mathbf{x}(t), \mathbf{X}(t), \mathbf{Z}(t)) | T \ge t, \{\mathbf{C}(v), v \le t\}, \{\mathbf{W}(v), v \le t\} \right] \\ = \; & \mathrm{E}_{\mathbf{x}(t)} \left[\lambda(t; \mathbf{x}(t), \mathbf{Z}(t)) | T \ge t, \{\mathbf{C}(v), v \le t\}, \{\mathbf{W}(v), v \le t\} \right], \end{aligned}$$

where the expectation $\mathrm{E}_{\mathbf{c}}$ is calculated over the true exposure histories $\{\mathbf{c}(v), v \le t\}$, and the expectation $\mathrm{E}_{\mathbf{x}(t)}$ is calculated over the true exposure $\mathbf{x}(t)$.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/harvardbiostat/paper198

Hence, for time t_k, conditional on $\{\mathbf{C}(t), t \le t_k\}$ and $\{\mathbf{W}(t), t \le t_k\}$, (1) induces

$$\begin{aligned} & \lambda(t_k; \{\mathbf{C}(t), t \le t_k\}, \{\mathbf{W}(t), t \le t_k\}) \\ = \; & \lambda_0(t_k) \exp(\boldsymbol{\beta}_z' \mathbf{Z}(t_k)) \cdot \mathrm{E}_{\mathbf{x}(t_k)}[\exp(\boldsymbol{\beta}_x' \mathbf{x}(t_k)) | \{\mathbf{C}(t), t \le t_k\}, \{\mathbf{W}(t), t \le t_k\}, T \ge t_k] \\ \approx \; & \lambda_0(t_k) \exp(\boldsymbol{\beta}_z' \mathbf{Z}(t_k)) \cdot \exp(\boldsymbol{\beta}_x' \cdot \mathrm{E}_{\mathbf{x}(t_k)}[\mathbf{x}(t_k) | \{\mathbf{C}(t), t \le t_k\}, \{\mathbf{W}(t), t \le t_k\}, T \ge t_k]), \quad (4) \end{aligned}$$

where the approximation in (4) is exact when, in each risk set, $(\mathbf{x}(t_k) | \{\mathbf{C}(t), t \le t_k\}$, $\{\mathbf{W}(t), t \le t_k\})$ is normal with homoscedastic covariance matrix. Otherwise, it will hold approximately when, in each risk set, $(\mathbf{x}(t_k) | \{\mathbf{C}(t), t \le t_k\}, \{\mathbf{W}(t), t \le t_k\})$ is normal and $\boldsymbol{\beta}_x' \Sigma_x(\mathbf{C}(t), \mathbf{W}(t)) \boldsymbol{\beta}_x$ is small, or, $\boldsymbol{\beta}_x$ is small with either homoscedastic covariance matrix or small $\boldsymbol{\beta}_x' \Sigma_x(\mathbf{C}(t), \mathbf{W}(t)) \boldsymbol{\beta}_x$ (Prentice, 1982; Liao et al., 2011).

When, as here, the $a - b$ moving average exposure is of interest, i.e., $\mathbf{x}(t_k) = \frac{1}{b-a+1} \cdot \sum_{j=k-b+1}^{k-a+1} \mathbf{c}(t_j)$, (4) can be written as

$$\begin{aligned} & \lambda(t_k; \{\mathbf{C}(t), t \le t_k\}, \{\mathbf{W}(t), t \le t_k\}) \\ \approx \; & \lambda_0(t_k) \exp(\boldsymbol{\beta}_z' \mathbf{Z}(t_k)) \cdot \exp(\overline{\boldsymbol{\beta}}_x' \cdot \mathrm{E}_{\overline{\mathbf{c}}_k}[\overline{\mathbf{c}}_k | \overline{\mathbf{C}}_k, \overline{\mathbf{W}}_k, T \ge t_k]), \end{aligned} \quad (5)$$

where $\overline{\mathbf{c}}_k = (\mathbf{c}(t_{k-b+1}), \cdots, \mathbf{c}(t_{k-a+1}))$, $\overline{\mathbf{C}}_k = (\mathbf{C}(t_{k-b+1}), \cdots, \mathbf{C}(t_k))$, $\overline{\mathbf{W}}_k = (\mathbf{W}(t_{k-b+1}),$ $\cdots, \mathbf{W}(t_k))$, $\overline{\boldsymbol{\beta}}_x = \frac{1}{b-a+1} \mathbf{1}_{b-a+1} \otimes \boldsymbol{\beta}_x$ with $\mathbf{1}_{b-a+1}' = (1, 1, \cdots, 1)_{1 \times (b-a+1)}$. For ease of notation, we assume that exposure is never missing over times $t_{k-b+1}, \cdots, t_{k-a+1}$, otherwise, missing exposure can be accommodated by (2). Note that $\overline{\mathbf{c}}_k = (\mathbf{c}(t_{k-b+1}), \cdots, \mathbf{c}(t_{k-a+1}))$ for the $a-b$ moving average, although the surrogate exposure histories are available up to time t_k and before time t_{k-b+1}. This is because $\mathbf{x}(t_k)$ doesn't depend on $(\mathbf{c}(t_{k-a+2}), \cdots, \mathbf{c}(t_k))$ for $a \ge 2$ and the integrals with regard to these variables are equal to one.

We assume a linear measurement error model for the point exposures at time t is given by

$$\mathrm{E}(\mathbf{c}(t) | \mathbf{C}(t), \mathbf{W}(t)) = \boldsymbol{\alpha}_0(t) + \boldsymbol{\alpha}_1'(t) \mathbf{C}(t) + \boldsymbol{\alpha}_2'(t) \mathbf{W}(t), \quad (6)$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

where $\dim(\boldsymbol{\alpha}_0(t)) = p \times 1$, $\dim(\boldsymbol{\alpha}_1(t)) = p \times p$, $\dim(\boldsymbol{\alpha}_2(t)) = q \times p$. Note that (6) implies the following localized error assumption (Zucker and Spiegelman, 2008) for any $j \leq k$ in a risk set:

$$\begin{aligned} \mathrm{E}(\mathbf{c}(t_j)|\{\mathbf{C}(t), t \leq t_k\}, \{\mathbf{W}(t), t \leq t_k\}) &= \mathrm{E}(\mathbf{c}(t_j)|\mathbf{C}(t_j), \mathbf{W}(t_j)) \\ &= \boldsymbol{\alpha}_0(t_j) + \boldsymbol{\alpha}_1'(t_j)\mathbf{C}(t_j) + \boldsymbol{\alpha}_2'(t_j)\mathbf{W}(t_j). \end{aligned} \tag{7}$$

This assumption will typically not be empirically verifiable in the main study/external validation study design because of the typical lack of repeatedly validated point exposures within subjects. Then, for the $a - b$ moving average exposure, the following set of measurement error models needs to be estimated in each risk set,

$$\mathrm{E}[\overline{\mathbf{c}}_k|\overline{\mathbf{C}}_k, \overline{\mathbf{W}}_k, T \geq t_k] = \begin{bmatrix} \boldsymbol{\alpha}_0(t_{k-b+1}) + \boldsymbol{\alpha}_1'(t_{k-b+1})\mathbf{C}(t_{k-b+1}) + \boldsymbol{\alpha}_2'(t_{k-b+1})\mathbf{W}(t_{k-b+1}) \\ \boldsymbol{\alpha}_0(t_{k-b+2}) + \boldsymbol{\alpha}_1'(t_{k-b+2})\mathbf{C}(t_{k-b+2}) + \boldsymbol{\alpha}_2'(t_{k-b+2})\mathbf{W}(t_{k-b+2}) \\ \cdots \\ \boldsymbol{\alpha}_0(t_{k-a+1}) + \boldsymbol{\alpha}_1'(t_{k-a+1})\mathbf{C}(t_{k-a+1}) + \boldsymbol{\alpha}_2'(t_{k-a+1})\mathbf{W}(t_{k-a+1}) \end{bmatrix}, \tag{8}$$

where $\boldsymbol{\alpha}_0(t_j)$, $\boldsymbol{\alpha}_1(t_j)$, $\boldsymbol{\alpha}_2(t_j)$, $j \in [k - b + 1, k - a + 1]$ are estimated in each risk set. Then, (5) can be written as

$$\lambda(t_k; \{\mathbf{C}(t), t \leq t_k\}, \{\mathbf{W}(t), t \leq t_k\}) \approx \lambda_0(t_k) \exp(\boldsymbol{\beta}_x'\hat{\mathbf{x}}(t_k) + \boldsymbol{\beta}_z'\mathbf{Z}(t_k)) \tag{9}$$

with $\hat{\mathbf{x}}(t_k) = \frac{1}{b-a+1}\sum_{j=k-b+1}^{k-a+1}[\hat{\boldsymbol{\alpha}}_0(t_j) + \hat{\boldsymbol{\alpha}}_1'(t_j)\mathbf{C}(t_j) + \hat{\boldsymbol{\alpha}}_2'(t_j)\mathbf{W}(t_j)]$. Once $\hat{\mathbf{x}}(t_k)$ are estimated in the main study using the measurement error model (6) fit in the validation study, the corrected point estimates $\hat{\boldsymbol{\beta}}_{corr}$ of $\boldsymbol{\beta} = (\boldsymbol{\beta}_x, \boldsymbol{\beta}_z)$ are obtained from fitting the ordinary Cox model (1) with $\hat{\mathbf{x}}(t_k)$.

The details of the algorithm is given below for the $a - b$ moving average exposure.

1. Order the r unique failure times that occur in the main study as $t_{k_1} < t_{k_2} < \cdots < t_{k_r}$,

CODRA
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/harvardbiostat/paper198

where $(k_1, k_2, \cdots, k_r)$ is a subset of indices on the entire time scale. Identify the subjects who belong to each of the r risk sets $R_m(t_{k_d})$, $d = 1, 2, \cdots, r$, in the main study. Generate the risk process indicator $Y_m(i, t_{k_d}) = I(T_i \geq t_{k_d})$.

2. Find the r risk sets $R_v(t_{k_d})$, $d = 1, 2, \cdots, r$, which consist of the subjects in the validation study who are at risk at time t_{k_d}. Generate the risk process indicator $Y_v(i, t_{k_d}) = I(T_i \geq t_{k_d})$. Estimate $\hat{\boldsymbol{\psi}}^{k_d}(t_j) = \{\hat{\boldsymbol{\alpha}}_{0k_d}(t_j), \hat{\boldsymbol{\alpha}}_{1k_d}(t_j), \hat{\boldsymbol{\alpha}}_{2k_d}(t_j)\}$, $j \in [k_d - b + 1, k_d - a + 1]$, by running the regression using the most recently observed exposures $\mathbf{c}^*(t_j)$ on $\mathbf{C}^*(t_j)$, $\mathbf{W}^*(t_j)$ prior to time t_j from the subjects in $R_v(t_{k_d})$. Since the validation data set is usually small, there will typically not be enough data to estimate the models for some t_j in the later risk sets. Hence, it will often to be necessary to estimate only $\hat{\boldsymbol{\psi}}^{k_d}(t_{k_d}) = \{\hat{\boldsymbol{\alpha}}_{0k_d}(t_{k_d}), \hat{\boldsymbol{\alpha}}_{1k_d}(t_{k_d}), \hat{\boldsymbol{\alpha}}_{2k_d}(t_{k_d})\}$, $d = 1, 2, \cdots, r$ in the validation study, approximating $\hat{\boldsymbol{\psi}}^{k_d}(t_j)$ by $\hat{\boldsymbol{\psi}}^{k_l}(t_{k_l})$, where $l \in (1, 2, \cdots, r)$ and t_{k_l} is the closest failure time to t_j. This approximation assumes, for a fixed t_j, $\boldsymbol{\psi}^{k_d}(t_j)$ are the same in all the risk sets $R_v(t_{k_e})$, where $k_e \geq k_d \geq j$.

3. Estimate $\hat{\mathbf{x}}_i(t_{k_d}) = \frac{1}{b-a+1} \sum_{j=k_d-b+1}^{k_d-a+1} [\hat{\boldsymbol{\alpha}}_{0k_d}(t_j) + \hat{\boldsymbol{\alpha}}'_{1k_d}(t_j)\mathbf{C}_i^*(t_j) + \hat{\boldsymbol{\alpha}}'_{2k_d}(t_j)\mathbf{W}_i^*(t_j)]$ for each subject i in each risk set $R_m(t_{k_d})$ in the main study, $d = 1, 2, ..., r$.

4. Fit the usual Cox model on $(\hat{\mathbf{x}}_i(t_{k_d}), \mathbf{Z}_i(t_{k_d}), Y_m(i, t_{k_d}), T_i, D_i)$ to get the measurement error corrected estimator $\hat{\boldsymbol{\beta}}_{corr}$, which solves

$$\sum_{i=1}^{n_1} \int_0^{t^*} \left\{ \begin{pmatrix} \hat{\mathbf{x}}_i(\hat{\boldsymbol{\psi}}, t) \\ \mathbf{Z}_i(t) \end{pmatrix} - \frac{\mathbf{S}^{(1)}(\boldsymbol{\beta}, \hat{\boldsymbol{\psi}}, t)}{S^{(0)}(\boldsymbol{\beta}, \hat{\boldsymbol{\psi}}, t)} \right\} N_i(\, dt) = 0, \tag{10}$$

where $S^{(0)}(\boldsymbol{\beta}, \hat{\boldsymbol{\psi}}, t) = {n_1}^{-1} \sum_{i=1}^{n_1} Y_m(i, t) \exp\{\boldsymbol{\beta}'_x \hat{\mathbf{x}}_i(\hat{\boldsymbol{\psi}}, t) + \boldsymbol{\beta}'_z \mathbf{Z}_i(t)\}$, and $\mathbf{S}^{(1)}(\boldsymbol{\beta}, \hat{\boldsymbol{\psi}}, t) = {n_1}^{-1} \sum_{i=1}^{n_1} Y_m(i, t) \begin{pmatrix} \hat{\mathbf{x}}_i(\hat{\boldsymbol{\psi}}, t) \\ \mathbf{Z}_i(t) \end{pmatrix} \exp\{\boldsymbol{\beta}'_x \hat{\mathbf{x}}_i(\hat{\boldsymbol{\psi}}, t) + \boldsymbol{\beta}'_z \mathbf{Z}_i(t)\}$, with the counting process $N_i(t) = I(T_i \leq t, D_i = 1)$.

COBRA BIOSTATISTICS WORKING PAPER Collection of Biostatistics Research Archive

Hosted by The Berkeley Electronic Press

Unlike the algorithm proposed in Liao et al. (2011), this algorithm constructs functions of the exposure histories from the point exposures, here the monthly point exposures, to utilize the limited information in the validation study as previously discussed.

3.2 Asymptotic Properties Of $\hat{\boldsymbol{\beta}}_{corr}$

The arguments for approximate consistency of $\hat{\boldsymbol{\beta}}_{corr}$ are similar to that given in Web Appendix B.1 of Liao et al. (2011) and will not be repeated here. The asymptotic normality of $\hat{\boldsymbol{\beta}}_{corr}$ follows similarly to the results in Web Appendix B.2 of Liao et al. (2011), that is, the asymptotic covariance of $\hat{\boldsymbol{\beta}}_{corr}$ is given by

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

where $\hat{\boldsymbol{I}}_{\boldsymbol{\beta}}^{-1} = \left[\frac{1}{n_1}\sum_{i=1}^{n_1}\frac{\partial \mathbf{U}_{\boldsymbol{\beta}i}(\boldsymbol{\beta}|\hat{\boldsymbol{\psi}})}{\partial \boldsymbol{\beta}}\right]^{-1}\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_{corr}}$, $\hat{\mathbf{H}}_{\boldsymbol{\beta},\boldsymbol{\psi}} = \hat{\mathbf{H}}_{\boldsymbol{\beta}} + \frac{1}{n_1 n_2}\hat{\mathbf{U}}^*(\boldsymbol{\beta},\boldsymbol{\psi})\hat{\mathbf{V}}_{\boldsymbol{\psi}}\hat{\mathbf{U}}^*(\boldsymbol{\beta},\boldsymbol{\psi})'$, $\hat{\mathbf{H}}_{\boldsymbol{\beta}} = \frac{1}{n_1}\sum_{i=1}^{n_1}\tilde{\mathbf{U}}_{\boldsymbol{\beta}i}(\boldsymbol{\beta},\hat{\boldsymbol{\psi}})\tilde{\mathbf{U}}'_{\boldsymbol{\beta}i}(\boldsymbol{\beta},\hat{\boldsymbol{\psi}})|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_{corr}}$, with

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

and here, $\mathbf{U}_{\boldsymbol{\beta}i}(\boldsymbol{\beta}|\hat{\boldsymbol{\psi}}) = D_i\left\{\begin{pmatrix}\hat{\mathbf{x}}_i(T_i)\\ \mathbf{Z}_i(T_i)\end{pmatrix} - \frac{\mathbf{S}^{(1)}(\boldsymbol{\beta},\hat{\boldsymbol{\psi}},T_i)}{S^{(0)}(\boldsymbol{\beta},\hat{\boldsymbol{\psi}},T_i)}\right\}$ is the score equation for each subject i, with $\hat{\boldsymbol{\psi}} = (\hat{\boldsymbol{\psi}}^{k_l}(t_j), l = 1,2,\cdots,r; j = k_l - b + 1,\cdots,k_l - a + 1)$, $\dim(\hat{\boldsymbol{\psi}}) = p \times (p+q+1)r\times(b-a+1)$ and $\dim(\mathbf{U}_{\boldsymbol{\beta}i}) = (p+q)\times 1$. $\hat{\mathbf{U}}^*(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\psi}}) = \sum_{i=1}^{n_1}\frac{\partial \mathbf{U}_{\boldsymbol{\beta}i}(\boldsymbol{\beta}|\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}|_{(\boldsymbol{\beta},\boldsymbol{\psi})=(\hat{\boldsymbol{\beta}}_{corr},\hat{\boldsymbol{\psi}})}$, with $\dim(\hat{\mathbf{U}}^*) = (p+q)\times p(p+q+1)r\times(b-a+1)$. $\text{Cov}(\hat{\boldsymbol{\psi}})$ can be estimated by $\frac{1}{n_2}\hat{\mathbf{V}}_{\hat{\boldsymbol{\psi}}}$, with $\dim(\mathbf{V}_{\hat{\boldsymbol{\psi}}}) = p(p+q+1)r(b-a+1)\times p(p+q+1)r(b-a+1)$ and $\hat{\mathbf{V}}_{\hat{\boldsymbol{\psi}}}$ is constructed

COBRA
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/harvardbiostat/paper198

as

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

where, in the validation study, $\hat{\boldsymbol{\psi}}$ solves $\sum_{i=1}^{n_2}\mathbf{U}_{\boldsymbol{\psi} i,k_d}(\boldsymbol{\psi})=\mathbf{0}$, with $\mathbf{U}_{\boldsymbol{\psi} i,k_d}(\boldsymbol{\psi}) = (\mathbf{U}'_{\boldsymbol{\alpha}_0 i,k_d}(\boldsymbol{\psi}),$ $\mathbf{U}'_{\boldsymbol{\alpha}_1 i,k_d}(\boldsymbol{\psi}),\ \mathbf{U}'_{\boldsymbol{\alpha}_2 i,k_d}(\boldsymbol{\psi}))$ is defined as follows

$$\begin{aligned}
\mathbf{U}_{\boldsymbol{\alpha}_0 i,k_d}(\boldsymbol{\psi}) &= Y_v(i,t_{k_d})[\mathbf{c}_i(t_j)-\boldsymbol{\alpha}_{0k_d}(t_j)-\boldsymbol{\alpha}_{1k_d}(t_j)\mathbf{C}_i(t_j)-\boldsymbol{\alpha}_{2k_d}(t_j)\mathbf{W}_i(t_j)]'\\
\mathbf{U}_{\boldsymbol{\alpha}_1 i,k_d}(\boldsymbol{\psi}) &= Y_v(i,t_{k_d})\mathbf{C}_i(t_j)[\mathbf{c}_i(t_j)-\boldsymbol{\alpha}_{0k_d}(t_j)-\boldsymbol{\alpha}_{1k_d}(t_j)\mathbf{C}_i(t_j)-\boldsymbol{\alpha}_{2k_d}(t_j)\mathbf{W}_i(t_j)]'\\
\mathbf{U}_{\boldsymbol{\alpha}_2 i,k_d}(\boldsymbol{\psi}) &= Y_v(i,t_{k_d})\mathbf{W}_i(t_j)[\mathbf{c}_i(t_j)-\boldsymbol{\alpha}_{0k_d}(t_j)-\boldsymbol{\alpha}_{1k_d}(t_j)\mathbf{C}_i(t_j)-\boldsymbol{\alpha}_{2k_d}(t_j)\mathbf{W}_i(t_j)]' \quad (12)
\end{aligned}$$

for $d=1,\cdots,r,\ i=1,\cdots,n_2,\ j=k_d-b+1,\cdots,k_d-a+1$, $\dim(\mathbf{U}_{\boldsymbol{\alpha}_0 i,k_d})=1\times p\times(b-a+1)$, $\dim(\mathbf{U}_{\boldsymbol{\alpha}_1 i,k_d})=p\times p\times(b-a+1)$, $\dim(\mathbf{U}_{\boldsymbol{\alpha}_2 i,k_d})=q\times p\times(b-a+1)$, $\dim(\mathbf{U}_{\boldsymbol{\psi} i,k_d}(\boldsymbol{\psi}))=p\times(p+q+1)\times(b-a+1)$, $\mathbf{U}_{\boldsymbol{\psi} i}=(\mathbf{U}_{\boldsymbol{\psi} i,k_1},\mathbf{U}_{\boldsymbol{\psi} i,k_2},\cdots,\mathbf{U}_{\boldsymbol{\psi} i,k_r})$ and $\dim(\mathbf{U}_{\boldsymbol{\psi} i})=p\times(p+q+1)r\times(b-a+1)$.

Since the risk sets in the validation study are small, the bias-corrected robust covariance estimator proposed by Mancl and DeRouen (2001) is used for $\hat{\mathbf{V}}_{\hat{\boldsymbol{\psi}}}$ to improve the performance of the variance estimator. It will be shown that the coverage probability was close to the nominal level of 95% in our simulation studies with this bias-corrected variance estimator in the next section.

4 SIMULATION STUDIES

We conducted finite sample simulation studies under scenarios arising within a main study/external validation study design. Since there is no other method that handles the data structure considered here, results for the new estimator are compared to the naive estimator. Motivated by our data, the NHS AP cohort and nine city validation study, estimation and inference for the 12-month moving average exposure with $a=1$ and

Hosted by The Berkeley Electronic Press

$b = 12$ was the focus. Following the motivating data, we set the study duration to $t^* = 84$ months (7 years) of follow-up. Both the rare disease and the common disease scenarios were considered. In the rare disease scenario, the cumulative incidence of the outcome was set at 1%, and the sample size for the main study and the validation study was set to $n_1 = 50000$ and $n_2 = 500$, respectively. In the common disease scenario, the sample size for the main study and the validation study was set to $n_1 = 1000$ and $n_2 = 500$, respectively, while the cumulative incidence was set to 50%, thereby fixing the number of cases to be the same in the two scenarios in order to make the results more comparable with one other. In addition, to more closely follow our data, we also studied a scenario with an 8% cumulative incidence and a sample size of $n_1 = 100000$ for the main study and $n_2 = 100$ for the validation study. All simulation results are based on 1000 replications, where the corresponding 95% nominal coverage is $[93.6\%, 96.4\%]$.

We first generated the surrogate exposure $\boldsymbol{C}_i \sim MVN(\boldsymbol{\mu}_C, \boldsymbol{\Sigma}_C)$, where $\boldsymbol{C}_i$ is a $m-$vector with m 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 $\Sigma_C(j,l) = 1$ if $j = l$ and $\Sigma_C(j,l) = \rho_I^{|j-l|^\tau}$ if $j \neq l$, with $\tau \in [0,1]$, and ρ_I as the intra-class correlation. Under this assumed damped exponential (DEX) covariance structure (Munoz et al., 1992), $\hat{\tau} = 0.30$ in the case study. When $\tau = 0$, a compound symmetry (CS) 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 of the intraclass correlation coefficients between the possible time points, $\bar{\rho}_{I_{AR}}$, over $[0, t^*]$ equal to $\rho_{I_{CS}}$, i.e., $\frac{1}{t^*}\int_0^{t^*}(\rho_{I_{AR}})^t dt = \rho_{I_{CS}}$. Solving this equation, when $t^* = 84$, we obtained values of $\rho_{I_{AR}}$ as 0.963, 0.987, 0.997, as would occur between two measurements adjacent in time, and at two measurements the farthest apart, when one corresponds to $t = 1$ and the other to $t = 84$, the correlation between these two would be 0.042, 0.329, and 0.809.

COBRA
BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/harvardbiostat/paper198

Next, the true exposure was generated as $c_{ij} = \alpha_0(j) + \alpha_1(j)C_{ij} + e_{ij}$, where $e_{ij} \sim N(0, \Delta_j)$, where Δ_j is the measurement error variance, characterized by $\Delta_j = \alpha_1^2(j)(\frac{1}{\rho^2} - 1)$, with $\rho = \text{Corr}(c_{ij}, C_{ij})$ describing the extent of measurement error between two point exposures taken at a single time within the same subject. $\alpha_0(j)$ and $\alpha_1(j)$ were set to step functions as $\alpha_0(j) = \begin{cases} 1.2, & \text{if } 1 \le j \le 40 \\ 1.5, & \text{if } 41 \le j \le 84 \end{cases}$, $\alpha_1(j) = \begin{cases} 0.5, & \text{if } 1 \le j \le 40 \\ 0.7, & \text{if } 41 \le j \le 84 \end{cases}$. In the main study, we generated pairs (c_{ij}, C_{ij}) at multiple time points t_j for each subject i. Because validation study subjects will only have one or two pairs of measurements, to have r pairs of measurements for each validation subject i, we first generated a vector $\mathbf{U}_i$ of 84 random numbers in $[0, 1]$ and sorted them. Then, the pairs of measurements with the r smallest elements of $\mathbf{U}_i$ were retained in the validation study for each subject i. For example, if $r = 2$, and the first two elements of $\mathbf{U}_i$ had the rank of 15 and 24 among the 84 random numbers generated, then $(c_i(15), C_i(15))$ and $(c_i(24), C_i(24))$ were included in the validation study for this subject.

The 12-month moving average exposure $x(t)$ and $X(t)$ was then generated as $x_i(t_k) = \frac{1}{12}\sum_{j=k-11}^{k} c_i(t_j)$, $X_i(t_k) = \frac{1}{12}\sum_{j=k-11}^{k} C_i(t_j)$ for $1 \le k \le m$ for the time points $\{t_0, t_1, t_2, \cdots, t_m\}$, $m = 84$. For simplicity, we used the cumulative average exposure for the first 12 months, then the 12-month moving average exposure for all 72 months following. Otherwise, as in the analysis to appear in the next section, we would need to exclude data from the first 12 months from the main study analysis, in order to be able to calculate a full 12 months of the 12-month moving average exposure for all study subjects from the start.

We generated the survival time data based on $x_i(t)$, following Web Appendix A.2 in Liao et al. (2011). Since each validation study subject has only a few measurements, the validation study risk sets can end up with small sample sizes. In addition to using a bias-corrected robust covariance estimator (Mancl and DeRouen, 2001) to improve finite sample performance in both the simulation study and the data analysis, we also grouped

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

the risk sets to ensure that they each had at least s observations, with $s = 10$ here, to further stabilize the results. This idea of grouping risk sets is similar to that used in the follow-up time regression calibration method proposed by Zhao and Prentice (2014).

Table 1 presents the results for the common and rare disease scenarios under AR(1) covariance, with 1 and 5 measurements per validation study subject. These simulations show that the corrected estimator had much smaller bias than the naive estimator in both point and interval estimates in all the scenarios we explored. Most of the corrected estimator's 95% coverage probabilities were within the nominal 95% coverage bounds of $[93.6\%, 96.4\%]$ for 1000 replications, while a few were slightly out of range. The simulation results under a CS covariance structure are presented in Supplementary Table 1 and showed the similar performance, with slightly less bias observed under AR(1) compared to CS for the same average correlation over the observed period. In Figure 1, we plotted the relative bias in $\hat{\beta}_{corr}$ by the validation study size under AR(1) covariance with $\bar{\rho}_{I_{AR}} = 0.6$. The analogous plot for CS covariance with $\rho_{I_{CS}} = 0.6$ is presented in Supplementary Figure 1. The new method had slightly less bias in the rare disease scenario (dashed line) than in the common disease scenario (solid line). It can be seen that bias was significantly reduced as the size of the validation study increased.

Because the damped exponential (DEX) covariance structure fit our data better than CS or AR(1) (Munoz et al., 1992), with $\hat{\rho}_I = 0.56$ and $\hat{\tau} = 0.30$, we conducted further simulations with this structure. To closely match the data motivating this research, the cumulative incidence rate was set to 8%, the main study sample size was $n_1 = 100000$, and each subject had 84 monthly measurements. We set $\alpha_0(j)$ and $\alpha_1(j)$ to $\alpha_0(j) = \begin{cases} 0.5, & \text{if} \quad 1 \leq j \leq 40 \\ 0.3, & \text{if} \quad 41 \leq j \leq 84 \end{cases}$, $\alpha_1(j) = \begin{cases} 0.2, & \text{if} \quad 1 \leq j \leq 40 \\ 0.5, & \text{if} \quad 41 \leq j \leq 84 \end{cases}$, similar to what had been observed in the nine city validation study. We studied several combinations of the validation study sample size (n_2) and the number of measurements per validation study subjects ($\mathcal{R}$) to explore how the size of the validation study and the number of repeated measurements

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/harvardbiostat/paper198

Table 1: Results for the simulation study of 12-month moving average exposure with a AR(1) covariance structure, for different intra-class correlations ($\rho_{I_{AR}}$) and different amounts of measurement error (ρ).

		Estimated $\hat{\beta}(\hat{SE}[\hat{\beta}])$		Relative Bias(%)		95% CI Coverage(%)	
$\bar{\rho}_{I_{AR}}$	ρ	Naive	Corrected	Naive	Corrected	Naive	Corrected
There is 1 measurement per validation study subject							
		$n_1 = 1000$,	$n_2 = 500$,	Common disease			
0.3	0.4	0.322(0.049)	0.449(0.119)	-35.5	-10.3	5.5	88.3
	0.6	0.326(0.050)	0.472(0.094)	-34.8	-5.7	5.9	93.9
	0.9	0.329(0.049)	0.485(0.076)	-34.2	-3.1	6.4	96.2
0.6	0.4	0.324(0.048)	0.457(0.121)	-35.3	-8.7	4.5	89.2
	0.6	0.328(0.048)	0.476(0.092)	-34.5	-4.9	5.0	94.9
	0.9	0.327(0.048)	0.483(0.074)	-34.6	-3.5	4.1	95.3
0.9	0.4	0.325(0.047)	0.459(0.118)	-34.9	-8.2	3.9	90.8
	0.6	0.326(0.047)	0.474(0.091)	-34.9	-5.3	3.6	92.9
	0.9	0.329(0.047)	0.485(0.073)	-34.2	-3.1	5.2	95.5
		$n_1 = 50000$,	$n_2 = 500$,	Rare disease			
0.3	0.4	0.348(0.047)	0.478(0.129)	-30.4	-4.4	9.2	92.6
	0.6	0.347(0.048)	0.495(0.097)	-30.6	-0.9	10.8	96.1
	0.9	0.344(0.048)	0.498(0.074)	-31.1	-0.5	9.0	96.1
0.6	0.4	0.342(0.045)	0.482(0.127)	-31.6	-3.6	5.8	93.4
	0.6	0.347(0.045)	0.499(0.095)	-30.5	-0.2	7.8	96.4
	0.9	0.347(0.046)	0.500(0.070)	-30.6	0.1	8.4	96.5
0.9	0.4	0.345(0.044)	0.481(0.128)	-31.0	-3.7	5.3	93.2
	0.6	0.348(0.044)	0.499(0.093)	-30.5	-0.3	6.7	96.0
	0.9	0.345(0.045)	0.498(0.069)	-31.0	-0.5	5.6	95.1
There are 5 measurements per validation study subject							
		$n_1 = 1000$,	$n_2 = 500$,	Common disease			
0.3	0.4	0.321(0.049)	0.469(0.082)	-35.7	-6.1	4.8	92.0
	0.6	0.325(0.049)	0.479(0.076)	-35.0	-4.3	5.9	94.7
	0.9	0.327(0.049)	0.482(0.073)	-34.6	-3.5	6.8	95.2
0.6	0.4	0.326(0.048)	0.478(0.080)	-34.7	-4.5	5.2	94.1
	0.6	0.326(0.048)	0.481(0.074)	-34.7	-3.7	4.1	95.1
	0.9	0.326(0.048)	0.481(0.071)	-34.9	-3.9	4.3	94.6
0.9	0.4	0.322(0.047)	0.471(0.078)	-35.5	-5.9	3.1	92.6
	0.6	0.326(0.047)	0.482(0.073)	-34.8	-3.7	3.5	94.9
	0.9	0.328(0.047)	0.485(0.070)	-34.4	-3.0	3.8	96.1
		$n_1 = 50000$,	$n_2 = 500$,	Rare disease			
0.3	0.4	0.346(0.047)	0.495(0.080)	-30.7	-1.0	10.6	95.0
	0.6	0.345(0.048)	0.497(0.073)	-31.0	-0.6	11.0	95.0
	0.9	0.346(0.048)	0.499(0.069)	-30.8	-0.2	9.9	95.6
0.6	0.4	0.347(0.045)	0.498(0.078)	-30.7	-0.4	7.5	95.6
	0.6	0.346(0.045)	0.499(0.070)	-30.8	-0.2	8.1	94.7
	0.9	0.345(0.046)	0.498(0.066)	-31.0	-0.3	7.3	95.7
0.9	0.4	0.346(0.044)	0.496(0.076)	-30.8	-0.9	6.5	95.9
	0.6	0.345(0.044)	0.497(0.069)	-31.1	-0.6	5.8	95.6
	0.9	0.346(0.045)	0.500(0.065)	-30.8	-0.1	6.2	94.8

True $\beta = 0.5$, the study duration $t^* = 84$, the number of simulation replications $B = 1000$.
In the common disease situation, the cumulative incidence was about 50% with $n_1 = 1000$.
In the rare disease situation, the cumulative incidence was about 1% with $n_1 = 50000$.
$\bar{\rho}_{I_{AR}} = \frac{1}{t^*}\int_0^{t^*}(\rho_{I_{AR}})^t dt$ with $t^* = 84$, $\rho_{I_{AR}} = 0.963, 0.987, 0.997$ respectively.

Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

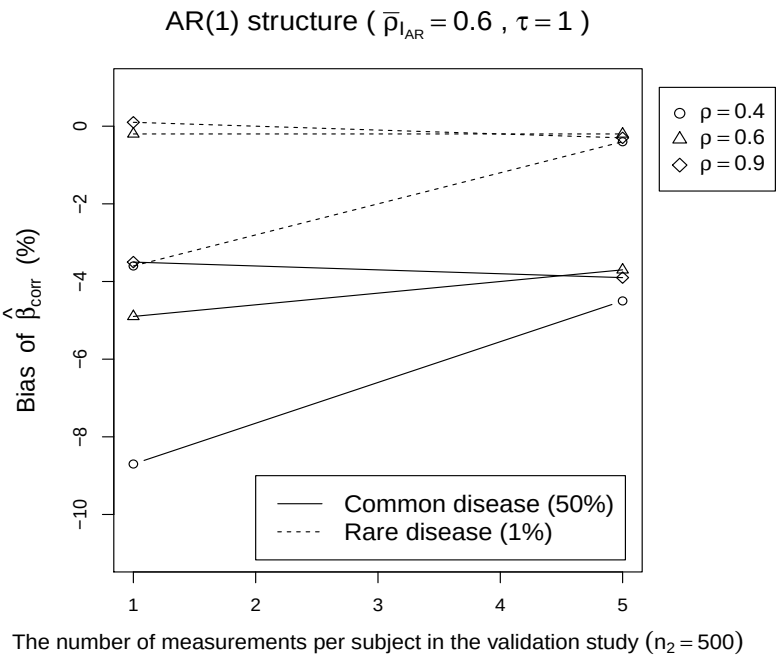

Figure 1: Plot for relative bias in relation to the validation study size for AR(1) covariance structure, $\beta = 0.5$.

for fixed validation study size affected the bias. The results are presented in Table 2 for several plausible correlations describing the extent of measurement error between the true exposure and the surrogate exposure, with excellent finite sample performance of the corrected estimator evident. The bias was less than 5% when the correlation between the true exposure and the surrogate exposure was greater than 0.6, as was the case for the spatio-temporal predicted $PM_{2.5}$ of ambient origin in relation to personal $PM_{2.5}$ exposure of ambient origin in the nine city validation study. When the correlation between the true exposure and the surrogate exposure was lower than 0.6, we saw somewhat more bias. As shown in Table 2, when the total person-time $n_2\mathcal{R}$ of the validation study was constant, for example, when $n_2\mathcal{R} = 200$, the bias was similar as n_2 and $\mathcal{R}$ were varied. The bias in both the point and interval estimates was significantly reduced when the total person-time in the validation study $n_2\mathcal{R}$ increased.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/harvardbiostat/paper198

Table 2: Results for the simulation study of 12-month moving average exposure with a DEX covariance structure, for intra-class correlation $\rho_{I_{DEX}} = 0.56$, $\tau = 0.30$. $n_1 = 100000$.

			Estimated $\hat{\beta}(\hat{SE}[\hat{\beta}])$		Percent Bias (%)		95% CI Coverage (%)	
n_2	$\mathcal{R}$	ρ	Naive	Corrected	Naive	Corrected	Naive	Corrected
		0.4	0.242(0.016)	0.428(0.158)	-51.6	-14.4	0.0	86.5
200	1	0.6	0.242(0.016)	0.475(0.105)	-51.5	-5.0	0.0	91.5
		0.9	0.241(0.016)	0.493(0.051)	-51.8	-1.3	0.0	95.5
		0.4	0.241(0.016)	0.425(0.158)	-51.7	-15.0	0.0	85.5
100	2	0.6	0.242(0.016)	0.476(0.105)	-51.6	-4.7	0.0	92.1
		0.9	0.242(0.016)	0.495(0.051)	-51.6	-0.9	0.0	96.0
		0.4	0.241(0.016)	0.422(0.156)	-51.8	-15.6	0.0	85.2
40	5	0.6	0.242(0.016)	0.474(0.102)	-51.5	-5.2	0.0	92.4
		0.9	0.242(0.016)	0.495(0.051)	-51.7	-0.9	0.0	95.8
		0.4	0.241(0.016)	0.428(0.160)	-51.7	-14.4	0.0	84.8
20	10	0.6	0.241(0.016)	0.476(0.106)	-51.8	-4.8	0.0	92.5
		0.9	0.241(0.016)	0.493(0.052)	-51.9	-1.4	0.0	95.2
		0.4	0.241(0.016)	0.458(0.103)	-51.7	-8.4	0.0	90.4
500	1	0.6	0.242(0.016)	0.488(0.072)	-51.6	-2.4	0.0	94.1
		0.9	0.243(0.016)	0.498(0.041)	-51.5	-0.3	0.0	96.1
		0.4	0.241(0.016)	0.451(0.105)	-51.8	-9.7	0.0	87.9
100	5	0.6	0.242(0.016)	0.482(0.074)	-51.6	-3.6	0.0	94.4
		0.9	0.242(0.016)	0.496(0.043)	-51.5	-0.7	0.0	95.9
		0.4	0.242(0.016)	0.468(0.077)	-51.7	-6.3	0.0	91.2
100	10	0.6	0.242(0.016)	0.490(0.056)	-51.6	-2.1	0.0	95.4
		0.9	0.241(0.016)	0.496(0.037)	-51.7	-0.9	0.0	94.3

True $\beta = 0.5$, the study duration $t^* = 84$, the number of simulation replications $B = 1000$.
n_2 is the number of subjects in the validation study.
$\mathcal{R}$ is the number of measurements per validation study subject.
The cumulative incidence was about 8%.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

5 DATA ANALYSIS

We applied the new method to the analysis of the NHS AP study described in Section 2 (Hart et al., 2015). While the method can be applied to any AP constituent exposures, including $PM_{2.5}$, PM_{10} and $PM_{2.5-10}$, and even multiple error-prone exposures in a single model where $p > 1$, the goal here was to estimate the prospective association between chronic $PM_{2.5}$ exposure and all-cause mortality, hence $p = 1$. The spatio-temporal predicted $PM_{2.5}$ (Paciorek et al., 2009; Yanosky et al., 2014) exposure was used as the surrogate exposure, although as noted previously, the method can be applied to adjust for bias due to measurement error when the nearest EPA monitor exposure is the surrogate, as well. The data from the nine city validation study, including personal $PM_{2.5}$ exposure of ambient origin, the gold standard, and the surrogate, the spatio-temporal predicted $PM_{2.5}$, were available and utilized to allow the measurement error correction. During the follow-up period of 2000 to 2006, 7,538,226 person-months were observed among 108,767 nurses, during which time 8,617 deaths occurred from all causes excluding accidents. The 12-month moving average exposure, the exposure metric of interest, was calculated at every person-month from the monthly spatio-temporal predicted $PM_{2.5}$. In the nine city validation study, there were 184 person-months observed among 100 subjects aged 54 and 87 years old. All of these person-months met the conditions that the pair of true exposure and surrogate exposure were available and that the age variable was not missing. Since we used age as the time scale in the Cox model, as is more suitable for prospective epidemiologic studies of chronic disease incidence and mortality (Korn et al., 1997), in this risk set regression calibration-type method, the age of subjects from the validation study needed to be matched with the age of subjects from the NHS AP cohort, the main study, to form the appropriate risk sets, hence younger subjects in the validation study were excluded from the analysis. Basic statistics for monthly $PM_{2.5}$ in the NHS AP cohort and the validation study are given in Table 3, together with the age distribution, indicating a

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/harvardbiostat/paper198

good concordance in both studies. The Pearson correlation between the spatio-temporal predicted $PM_{2.5}$ and personal $PM_{2.5}$ exposure from ambient origin was 0.65, and the Spearman correlation was 0.67, indicating a moderate amount of measurement error in the point exposures.

Table 3: Basic statistics of $PM_{2.5}$ exposure in the NHS AP cohort and the validation study used in our analysis.

Variable	Nurses' Health Study ($n_1 = 108,767$) (Person-month=7,538,226)		Nine city validation study ($n_2 = 100$) (Person-month=184)	
	Mean(SD) or %	(min, max)	Mean(SD) or %	(min, max)
Monthly personal $PM_{2.5}$ of ambient origin ($\mu g/m^3$)	-	-	10.1(4.4)	(2.6, 22.8)
Monthly spatio-temporal predicted $PM_{2.5}$ ($\mu g/m^3$)	11.9 (4.1)	(0.79, 39.8)	16.1(4.6)	(6.6, 24.4)
Age (in years)	69.0 (7.3)	(53.5, 87.3)	70.8 (7.8)	(54.0, 85.4)

We first estimated the hazard ratio in a basic model adjusted for a limited set of potential confounders: calendar year, season and region, where season was grouped into two categories: October-March as the heating season and April-September as the cooling season, and U.S. census region was grouped into four levels — Northeast, Midwest, West, and South. Then, we fit a multivariate model that included a much larger set of potential confounders, including smoking status (current/former/never), pack-years (number of packs/day multiplied by number of years of cigarette smoking), family history of MI (yes or no), BMI (kg/m^2), hypercholesterolemia (yes or no), median family income in census tract of residence (in \$1,000), median house value in census tract of residence (in \$1,000), physical activity (< 3, 3 to < 9, 9 to < 18, 18 to < 27, ≥ 27 MET hr/week), race (caucasian v.s. not), Alternate Healthy Eating Index (AHEI) (Chiuve et al., 2008), individual level socioeconomic status (nurses education level (RN degree v.s. not), parental occupations (housewife mother at age 16 v.s. not; professional or manager father at age 16 v.s. not), marital status (married v.s. not), and husband's education (less than high school, high school, greater than high school)). The most recent measurement of the full

COBRA
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

covariate history was used except for pack-years which was entered as a cumulative total exposure.

Kioumourtzoglou et al. (2014) reported significant heterogeneity by season in the measurement error model fit to the nine city validation study, requiring us to include an interaction term between season and $PM_{2.5}$ in the measurement error model used in the analysis, as follows:

$$\mathrm{E}(c(t)|C(t),\mathbf{W}(t)) = \alpha_0(t)+\alpha_1(t)*C(t)+\alpha_2(t)*W_1(t)+\alpha_3(t)*W_2(t)+\alpha_4(t)*C(t)*W_2(t), \quad (13)$$

where $c(t)$ represents personal $PM_{2.5}$ exposure of ambient origin, $C(t)$ represents spatio-temporal predicted $PM_{2.5}$, $W_1(t)$ represents region, $W_2(t)$ represents season. The method readily accommodates functions of $C(t)$ and $\mathbf{W}(t)$, for example, $C(t) * W_2(t)$ in the measurement error model (6), since the derivation in (5) still holds when such functions are included. Since the additional covariates in the multivariate model were only available in the NHS AP cohort, we have assumed that they are not associated with the measurement error model conditional on the covariates already included and thus they are not considered for entry in (13).

The results of the analysis are given in Table 4. The uncorrected RR (95% CI) for $\mathrm{PM}_{2.5}$ measured by the spatio-temporal predicted $PM_{2.5}$ exposure was 1.20(1.11,1.29) per $10\mu\mathrm{g}/m^3$ in the basic model, and 1.13(1.05,1.22) in the multivariate model, demonstrating some confounding by some of the additional covariates. Since the validation study was small ($n_2 = 100$ with 184 person-month observations), we grouped the risk sets to stabilize the results, to ensure that, as described in Section 4, each risk set would have a minimum of s observations. Different choice of s results in different compositions and numbers of risk sets. To determine the optimal value of s, we conducted a 20-fold cross-validation approach in the validation study, and calculated the mean square error, mean absolute error, and Pearson and Spearman correlations of the prediction error, for each choice of

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/harvardbiostat/paper198

s ranging from 10 to 100. The mean square error and the mean absolute error were the smallest, and the correlations were the biggest, when $s = 85$, resulting in 2 risk sets, where the first risk set had 86 person months with age ranging from 53.8 to 70.5 years, and the second had 98 person months with age ranging from 70.6 to 87.0 years. We proceeded with the analysis with this optimized s.

As always, we must assess whether the assumptions necessary for valid application of the method are satisfied. In particular, at least one of the sets of conditions following equation (4) must be satisfied. One of these sets of conditions is that the residuals of the measurement error models for each risk set as given by equation (6) are normal and $\beta_x^2\sigma^2$ is small, where σ^2 is the residual variance of the model (13) in each risk set. The diagnostic plots for normality are displayed in Figure 2, with the left column of the plots corresponding to the first risk set and the right column of the plots corresponding to the second risk set. The residual variance is around 10.0 in both risk sets, and if $\beta_x = 0.165$ as estimated from the multivariate model in Table 4, then $\beta_x^2\sigma^2 = 0.27 < 0.5$. The cutoff 0.5 was suggested by Kuha (1994) as the maximum value of $\beta_x^2\sigma^2$ allowed for valid application of regression calibration. The normality plots and histograms of the residuals suggest close conformation to the normal distribution, with the p-value for the Shapiro-Wilk test of normality being 0.23 for the younger risk set and 0.17 for the older risk set. Given normality and small $\beta_x^2\sigma^2$, the approximation needed in (4) will hold.

The corrected estimates were then calculated using our new method, and the results are presented in Table 4, together with the uncorrected estimates. In the basic model, the corrected RR (95% CI) per $10\mu g/m^3$ $PM_{2.5}$ was de-attenuated to 1.27(1.08,1.48) from 1.20(1.11,1.29), $p = 0.003$. The de-attenuation was similar for the multivariate model, in which the corrected RR (95% CI) per $10\mu g/m^3$ $PM_{2.5}$ was increased to 1.18(1.02,1.36) from the uncorrected 1.13(1.05,1.22). These results demonstrate that the risk of all-cause mortality from chronic exposure to $PM_{2.5}$ of ambient origin has been under-estimated

CODRA
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

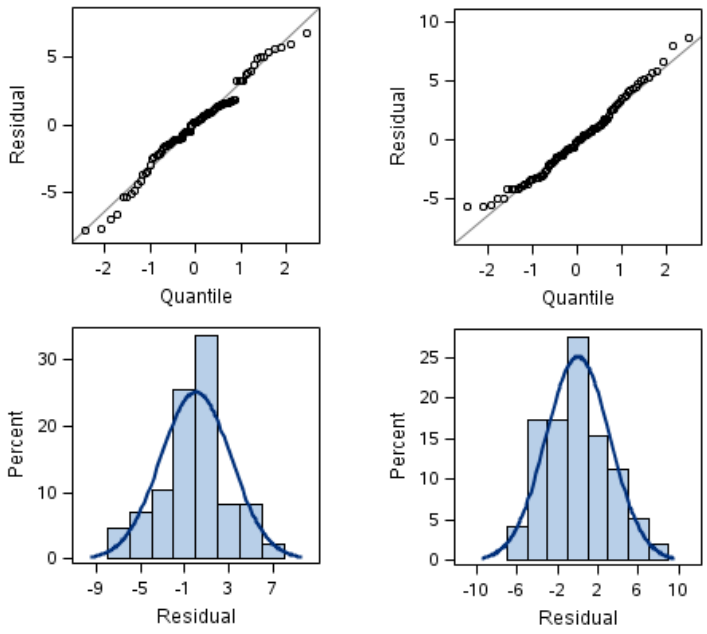

Figure 2: Normality plots for the residuals from the measurement error model (13). Left: Risk set 1; Right: Risk set 2.

when the spatio-temporal predicted exposure is used as the surrogate in the main study, and that our method is able to produce point and interval estimates of relative risk even with the limited validation data available. In the NHS AP cohort, the mean of the 12-month moving average $PM_{2.5}$ exposure was $12.0\mu g/m^3$ (Hart et al., 2015). Thus, this analysis provides evidence that $PM_{2.5}$ exposure below the current U.S. EPA standard of $12\mu g/m^3$ is associated with an increased risk of all-cause mortality, and that measurement error correction should be implemented in studies whenever possible to obtain valid point and interval estimate.

6 CONCLUSION AND DISCUSSION

Motivated by the NHS AP study, one of a small number of large scale prospective cohorts available worldwide in which chronic AP effects can be assessed, we have developed a RRC-type measurement error correction method for survival data analysis to obtain a consistent estimate of the hazard ratio for functions of mis-measured time-varying expo-

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/harvardbiostat/paper198

Table 4: Hazard ratios for $PM_{2.5}$ (per $10\mu g/m^3$) in relation to all-cause mortality in the NHS AP cohort.

	Basic model[a]			Multivariate model[b]		
Method	Estimate (S.E.)	RR (95% C.I.)	p-value	Estimate (S.E.)	RR (95% C.I.)	p-value
Uncorrected	0.181(0.039)	1.20(1.11,1.29)	< 0.001	0.125(0.039)	1.13(1.05,1.22)	0.002
Corrected	0.236(0.081)	1.27(1.08,1.48)	0.003	0.165(0.073)	1.18(1.02,1.36)	0.024

[a]Adjusted for age (in months), calendar year, region, and season.

[b]Adjusted for age (in months), calendar year, region, season, smoking status, pack-years, family history of MI, BMI, hypercholesterolemia, median family income in census tract of residence, median house value in census tract of residence, physical activity, race, Alternate Healthy Eating Index (AHEI), individual level socioeconomic status (nurses education level, parental occupations, marital status, and husband's education).

sure histories in the presence of limited validation data. Although motivated by studies of the health effects of chronic AP, the statistical methods developed in this paper have far wider applicability, as many longitudinal cohort studies in epidemiology focus on survival data endpoints such as cancer incidence and cause-specific mortality, in relation to exposures that are functions of mis-measured time-varying exposure histories. This fully general case arises, in addition to in AP epidemiology, e.g. Thomas et al. (1993), in occupational epidemiology, including studies of radiation exposure, e.g. Samet et al. (1991), and widely throughout nutritional epidemiology, e.g. Hu et al. (1999). The derivation given in (4) and (5) applies to all linear exposure metrics, including, importantly, the cumulative average exposure, defined as $\mathbf{x}(t_k) = \frac{1}{t_k - t_0} \sum_{j=0}^{k} \mathbf{c}(t_j)$, commonly used in nutritional epidemiology, and the cumulative total exposure, defined as $\mathbf{x}(t_k) = \sum_{j=0}^{k} \mathbf{c}(t_j)$, commonly used in occupational epidemiology. When either of these exposure functions is used, as given by equation (4), equation (5) will follow with $\overline{\mathbf{c}}_k = (\mathbf{c}(t_0), \cdots, \mathbf{c}(t_k))$, $\overline{\mathbf{C}}_k = (\mathbf{C}(t_0), \cdots, \mathbf{C}(t_k))$, $\overline{\mathbf{W}}_k = (\mathbf{W}(t_0), \cdots, \mathbf{W}(t_k))$, $\overline{\boldsymbol{\beta}}_x = \frac{1}{t_k - t_0} \mathbf{1}_{k+1} \otimes \boldsymbol{\beta}_x$ for the cumulative average exposure and $\overline{\boldsymbol{\beta}}_x = \mathbf{1}_{k+1} \otimes \boldsymbol{\beta}_x$ for the cumulative total exposure with $\mathbf{1}'_{k+1} = (1, 1, \cdots, 1)_{1 \times (k+1)}$. Then, in the algorithm given in Section 3.1, the range of the index j in Step 2 and the functional form in Step 3 can be altered as needed for the cumulative average exposure and the cumulative total exposure. The asymptotic variance in Section 3.2 will automatically apply with a different functional form for $\hat{\mathbf{x}}(t)$.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

This paper focused on the $a-b$ month moving average exposure as motivated by our case study, where the positive integers a and b are fixed constants. More research is being conducted for estimating the latency parameters, a and b, for different time-varying exposure metrics while simultaneously quantifying the hazard ratio for exposure in relation to health outcome (Wang et al., 2015). For example, in the $a-b$ month moving average exposure, instead of pre-specifying a and b as constants, a and b can be estimated together with the hazard ratio, $\boldsymbol{\beta}_x$, for different exposure metrics with and without exposure measurement error.

In the NHS AP cohort analysis, the age range of nurses was 33.8 years (406 months), during which time 367 unique failure times occured, with age(months) as the time scale. The age range of the subjects in the validation study was 31.4 years (377 months). Thus, we had at least one case occurring at nearly every age(months) of follow-up. For any time t_j on the observed time scale, we approximated $\hat{\boldsymbol{\psi}}^{k_d}(t_j)$ by $\hat{\boldsymbol{\psi}}^{k_l}(t_{k_l})$ for $d \geq l \in \{1, 2, \cdots, r\}$, where t_{k_l} is the closest failure time to t_j. In addition, we assumed that $\hat{\boldsymbol{\psi}}^{k_d}(t_j)$ was the same in all risk sets $R_v(t_{k_e})$, where $k_e \geq k_d \geq j$, using the largest risk set closest to t_j to calculate $\hat{\boldsymbol{\psi}}^{k_d}(t_j)$ for $t_{k_d} > t_j$. This approximation becomes better when the disease is rarer, which explained the slightly better performance of the rare disease scenario over the common disease scenario in our simulation studies. If more validation data become available in the future, it will be possible to fine-tune the validation study parameter estimation and better ensure the robustness of the estimator.

This is the first method developed for estimation and inference in survival data analysis of the effects of functions of mis-measured exposure histories, a common scenario in environmental, occupational, and nutritional epidemiology. The asymptotic variance estimator was derived. As shown in Xie et al. (2001) and Liao et al. (2011), some asymptotic bias in the estimator was expected due to the approximation in equation (4). Even, for example, when the data for $(\mathbf{x}(t_k)|\{\mathbf{C}(t), t \leq t_k\}, \{\mathbf{W}(t), t \leq t_k\})$ are generated from the

CORA
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/harvardbiostat/paper198

multivariate normal distribution at baseline such as in our simulations, the distribution would be expected to deviate from normality in later risk sets, inducing some degree of asymptotic bias. A detailed simulation study demonstrated improvements in the performance of the method as the validation study size increased. We suggest grouping the validation study risk sets so that the number of observations in each risk set will not be too small to stabilize the results, and proposed a data-driven cross-validation method to optimize the risk set size. When possible, a larger validation study with three or more measurements per subject would be desirable to further reduce bias, although the simulations indicated the results were reasonable for currently available validation study sizes.

For outcomes other than mortality, for example, lung cancer incidence, $\{T_i, D_i\}$ may be unknown in an external validation study, and the rare disease assumption will need to be invoked, setting $D_i = 0$. While our method is designed for the main study/external validation study design as motivated by our case study, it can be easily extended to the main study/internal validation study design (Zucker and Spiegelman, 2008), where the validation study is a sub-sample of the main study and thus the outcome status is also available in the validation study together with both true and surrogate exposures. Technical details for this extension will be covered in a future paper on measurement error correction analysis for the cumulative exposure to radon in relation to lung cancer mortality among uranium miners, updating Samet et al. (1991).

A user-friendly SAS macro has been developed and posted at `https://www.hsph.harvard.edu/donna-spiegelman/software/rrc-macro/`. The macro supports five commonly used exposure metrics which arise in environmental, occupational and nutritional epidemiology: the $a-b$ time unit moving average exposure, as is our focus here, the cumulative average exposure, the cumulative total exposure, the simple updated exposure (i.e., time-varying point exposure), and the time-independent exposure. In order to facilitate

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

valid estimation for U.S. AP studies, the nine city validation study has been posted at http://www.hsph.harvard.edu/pm2-5-validation-dataset/download-validation-dataset/.

7 SUPPLEMENTARY MATERIAL

Supplementary Table 1 : Results for the simulation study of 12-month moving average exposure with a compound symmetry covariance structure, for different intra-class correlations ($\rho_{I_{CS}}$) and different amounts of measurement error (ρ).

Supplementary Figure 1 : Plot for relative bias in relation to the validation study size for CS covariance structure.

References

Atkinson, R. W., Kang, S., Anderson, H. R., Mills, I. C., and Walton, H. A. (2014). Epidemiological time series studies of $PM_{2.5}$ and daily mortality and hospital admissions: a systematic review and meta-analysis. *Thorax*, 69(7):660–665.

Atkinson, R. W., Mills, I. C., Walton, H. A., and Anderson, H. R. (2015). Fine particle components and health – a systematic review and meta-analysis of epidemiological time series studies of daily mortality and hospital admissions. *J Expos Sci Environ Epidemiol*, 25(2):208–214.

Brook, R. D., Franklin, B., Cascio, W., Hong, Y., Howard, G., Lipsett, M., Luepker, R., Mittleman, M., Samet, J., Smith, S. C. J., and Tager, I. (2004). Air pollution and cardiovascular disease: a statement for healthcare professionals from the expert panel on population and prevention science of the american heart association. *Circulation*, 109(21):2655–2671.

COBRA A BEPRESS REPOSITORY Collection of Biostatistics Research Archive

http://biostats.bepress.com/harvardbiostat/paper198

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition*. Chapman and Hall.

Chakrabarti, B., Fine, P. M., Delfino, R., and Sioutas, C. (2004). Performance evaluation of the active-flow personal dataram $PM_{2.5}$ mass monitor (Thermo Anderson pDR-1200) designed for continuous personal exposure measurements. *Atmospheric Environment*, 38(20):3329–3340.

Chiuve, S. E., Rexrode, K. M., Spiegelman, D., Logroscino, G., Manson, J. E., and Rimm, E. B. (2008). Primary prevention of stroke by healthy lifestyle. *Circulation*, 118(9):947–54.

Colditz, G. A. and Hankinson, S. E. (2005). The Nurses' Health Study: lifestyle and health among women. *Nat Rev Cancer*, 5(5):388–396.

Colditz, G. A., Martin, P., Stampfer, M. J., Willett, W. C., Sampson, L., Rosner, B., Hennekens, C. H., and Speizer, F. E. (1986). Validation of questionnaire information on risk factors and disease outcomes in a prospective cohort study of women. *Am J Epidemiol*, 123(5):894–900.

Dab, W., Segala, C., Dor, F., Festy, B., Lameloise, P., Le Moullec, Y., Le Tertre, A., Medina, S., Quenel, P., Wallaert, B., and Zmirou, D. (2001). Air pollution and health: correlation or causality? the case of the relationship between exposure to particles and cardiopulmonary mortality. *J Air Waste Manag Assoc*, 51(2):220–235.

Hart, J. E., Liao, X., Hong, B., Puett, R. C., Yanosky, J. D., Suh, H., Kioumourtzoglou, M.-A., Spiegelman, D., and Laden, F. (2015). The association of long-term exposure to $PM_{2.5}$ on all-cause mortality in the Nurses' Health Study and the impact of measurement-error correction. *Environ Health*, 14(1):38.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Hoek, G., Krishnan, R. M., Beelen, R., Peters, A., Ostro, B., Brunekreef, B., and Kaufman, J. D. (2013). Long-term air pollution exposure and cardio- respiratory mortality: a review. *Environ Health*, 12(1):43.

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

Kioumourtzoglou, M.-A., Spiegelman, D., Szpiro, A. A., Sheppard, L., Kaufman, J. D., Yanosky, J. D., Williams, R., Laden, F., Hong, B., and Suh, H. (2014). Exposure measurement error in $PM_{2.5}$ health effects studies: a pooled analysis of eight personal exposure validation studies. *Environ Health*, 13(1):2.

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. *Am J Epidemiol*, 145(1):72–80.

Kuha, J. (1994). Corrections for exposure measurement error in logistic regression models with an application to nutritional data. *Stat Med*, 13(11):1135–48.

Lee, S. J., Demokritou, P., Koutrakis, P., and Delgado-Saborit, J. M. (2006). Development and evaluation of personal respirable particulate sampler (PRPS). *Atmospheric Environment*, 40(2):212–224.

Liao, X. M., Zucker, D. M., Li, Y., and Spiegelman, D. (2011). Survival analysis with error-prone time-varying covariates: a risk set calibration approach. *Biometrics*, 67(1):50–58.

Mancl, L. A. and DeRouen, T. A. (2001). A covariance estimator for gee with improved small-sample properties. *Biometrics*, 57(1):126–134.

Munoz, A., Carey, V., Schouten, J. P., Segal, M., and Rosner, B. A. (1992). A parametric

COBRA
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/harvardbiostat/paper198

family of correlation structures for the analysis of longitudinal data. *Biometrics*, 48:733–742.

Paciorek, C., Yanosky, J., Puett, R., Laden, F., and Suh, H. (2009). Practical large-scale spatio-temporal modeling of particulate matter concentrations. *Ann Appl Stat*, 3(1):370–397.

Pope III, C. A. (2000). Epidemiology of fine particulate air pollution and human health: biologic mechanisms and who's at risk? *Environmental Health Perspectives*, 108(Suppl 4):713–723.

Pope III, C. A. and Dockery, D. W. (2006). Health effects of fine particulate air pollution: lines that connect. *J Air Waste Manag Assoc*, 56(6):709–42.

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

Puett, R. C., Schwartz, J., Hart, J. E., Yanosky, J. D., Speizer, F. E., Suh, H., Paciorek, C. J., Neas, L. M., and Laden, F. (2008). Chronic particulate exposure, mortality, and coronary heart disease in the Nurses' Health Study. *Am J Epidemiol*, 168(10):1161–1168.

Rosner, B. A., 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. *Am J Epidemiol*, 132(4):734–45.

Rosner, B. A., Willett, W. C., and Spiegelman, D. (1989). Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Stat Med*, 8(9):1051–69; discussion 1071–3.

Samet, J. M., Pathak, D. R., Morgan, M. V., Key, C. R., Valdivia, A. A., and Lubin,

COBRA
Collection of Biostatistics

Hosted by The Berkeley Electronic Press

J. H. (1991). Lung cancer mortality and exposure to radon progeny in a cohort of new mexico underground uranium miners. *Health Phys*, 61(6):745–752.

Sarnat, J. A., Long, C. M., Koutrakis, P., Coull, B. A., Schwartz, J., and Suh, H. H. (2002). Using sulfur as a tracer of outdoor fine particulate matter. *Environ Sci Technol*, 36(24):5305–14.

Schwartz, J., Coull, B., Laden, F., and Ryan, L. (2008). The effect of dose and timing of dose on the association between airborne particles and survival. *Environ Health Perspect*, 116(1):64–69.

Sheppard, L., Slaughter, J. C., Schildcrout, J., Liu, L.-J. S., and Lumley, T. (2005). Exposure and measurement contributions to estimates of acute air pollution effects. *J Expo Anal Environ Epidemiol*, 15(4):366–76.

Spiegelman, D., McDermott, A., and Rosner, B. A. (1997). Regression calibration method for correcting measurement-error bias in nutritional epidemiology. *Am J Clin Nutr*, 65(4 Suppl):1179S–1186S.

Thomas, D., Stram, D., and Dwyer, J. (1993). Exposure measurement error: influence on exposure-disease relationships and methods of correction. *Annual Review of Public Health*, 14(1):69–93.

U.S. EPA. (2004). Air quality criteria for particulate matter (Final Report, Oct 2004). Technical Report EPA 600/P-99/002aF-bF, U.S. Environmental Protection Agency, Washington, DC.

Wang, C. Y., Hsu, L., Feng, Z. D., and Prentice, R. L. (1997). Regression calibration in failure time regression. *Biometrics*, 53(1):131–45.

Wang, M., Liao, X., Laden, F., and Spiegelman, D. (2015). Quantifying risk over the life

COBRA
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/harvardbiostat/paper198

course – latency, age-related susceptibility, and other time-varying exposure metrics. Under review.

Wilson, W. E., Mage, D. T., and Grant, L. D. (2000). Estimating separately personal exposure to ambient and nonambient particulate matter for epidemiology and risk assessment: why and how. *J Air Waste Manag Assoc*, 50(7):1167–83.

Xie, S. X., Wang, C. Y., and Prentice, R. L. (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.

Yanosky, J. D., Paciorek, C. J., Laden, F., Hart, J. E., Puett, R. C., Liao, D., and Suh, H. H. (2014). Spatio-temporal modeling of particulate air pollution in the conterminous united states using geographic and meteorological predictors. *Environ Health*, 13:63.

Zhao, S. and Prentice, R. L. (2014). Covariate measurement error correction methods in mediation analysis with failure time data. *Biometrics*, 70(4):835–844.

Zucker, D. M. and Spiegelman, D. (2008). Corrected score estimation in the proportional hazards model with misclassified discrete covariates. *Stat Med*, 27(11):1911–1933.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press