Environmental Research 134 (2014) 482–487

Contents lists available at ScienceDirect

Environmental Research

journal homepage: www.elsevier.com/locate/envres

Indirect adjustment for multiple missing variables applicable to environmental epidemiology

Hwashin H. Shin [a,b], Sabit Cakmak [a], Orly Brion [a], Paul Villeneuve [a,c], Michelle C. Turner [d], Mark S. Goldberg [e], Michael Jerrett [f], Hong Chen [g], Dan Crouse [a], Paul Peters [h], C Arden Pope III[i], Richard T. Burnett [a,*]

[a] Population Studies Division, Health Canada, Ottawa, Canada
[b] Department of Mathematics and Statistics, Queen's University, Kingston, Canada
[c] Division of Occupational and Environmental Health, Dalla Lama School of Public Health, University of Toronto, Toronto, Canada
[d] Institute of Population Health, University of Ottawa, Ottawa, Canada
[e] Department of Medicine, McGill University, Montreal, Canada
[f] School of Public Health, University of California, Berkeley, CA, USA
[g] Public Health Ontario, Toronto, Ontario, Canada
[h] Statistics Canada, Ottawa, Canada
[i] Department of Economics, Brigham Young University, Provo, USA

ARTICLE INFO

Available online 24 June 2014

Keywords:
Indirect adjustment
Cohort study
Air pollution
Survival analysis
Simulation

ABSTRACT

Objectives: Develop statistical methods for survival models to indirectly adjust hazard ratios of environmental exposures for missing risk factors.
Methods: A partitioned regression approach for linear models is applied to time to event survival analyses of cohort study data. Information on the correlation between observed and missing risk factors is obtained from ancillary data sources such as national health surveys. The relationship between the missing risk factors and survival is obtained from previously published studies. We first evaluated the methodology using simulations, by considering the Weibull survival distribution for a proportional hazards regression model with varied baseline functions, correlations between an adjusted variable and an adjustment variable as well as selected censoring rates. Then we illustrate the method in a large, representative Canadian cohort of the association between concentrations of ambient fine particulate matter and mortality from ischemic heart disease.
Results: Indirect adjustment for cigarette smoking habits and obesity increased the fine particulate matter-ischemic heart disease association by 3%–123%, depending on the number of variables considered in the adjustment model due to the negative correlation between these two risk factors and ambient air pollution concentrations in Canada. The simulations suggested that the method yielded small relative bias (< 40%) for most cohort designs encountered in environmental epidemiology.
Conclusions: This method can accommodate adjustment for multiple missing risk factors simultaneously while accounting for the associations between observed and missing risk factors and between missing risk factors and health endpoints.

1. Introduction

The issue of bias from omitted variables that may confound an association between a given outcome and exposure has been of interest in occupational epidemiology for many years. The main concern with many of these studies was that the sampling frame often comprised records that did not include data on personal risk factors, such as cigarette smoking. The nested case-control design and case-cohort study are approaches that were developed to address this challenge, with additional data on essential risk factors gathered from a subset of the cohort, thereby reducing costs considerably (Liddel et al., 1977; Langholz and Goldstein, 1996). Another approach to account for unmeasured confounding involves partitioning the incidence rate into components representing the exposure and confounding variables, thereby allowing for an indirect adjustment (Axelson, 1980). This method was developed for the case of incidence rates of disease in relation to a dichotomous exposure for a single risk factor, such as never/ever

[*] Correspondence to: Population Studies Division, Health Canada, 50 Columbine Driveway, Room 134, Ottawa, Ontario, Canada K1A 0K9. Fax: +1 613 941 3883.
E-mail address: rick.burnett@hc-sc.gc.ca (R.T. Burnett).

http://dx.doi.org/10.1016/j.envres.2014.05.016

smoking cigarettes. This indirect adjustment approach was augmented to estimate variances on the corrected rate ratios using Monte Carlo simulations (Steenland and Greenland, 2004) and it was further extended to account for an unmeasured continuous exposure variable for a single categorical risk factor (Villeneuve et al., 2010). These indirect methods are limited because confounding is not usually restricted to a single categorical risk factor but to several accepted risk factors that can take on several possible functional forms.

We have encountered recently a similar problem of unmeasured risk factors in conducting cohort studies of air pollution and health. In this paper we illustrate a new method using a cohort study that is a representative sample of the Canadian population. The study makes use of a random sample of citizens who completed the 1991 Canadian census long-form and who were subsequently followed-up in time to ascertain vital status and underlying cause of death through a probabilistic record linkage to the Canadian National Mortality Database up to 2001 (Wilkins et al., 2008). We then linked estimates of ambient fine particulate air pollution to the home address 6-digit postal code available in the 1991 census (van Donkelaar et al., 2010). Although some information on known risk factors for mortality was available, such as income, education, and occupation, other essential risk factors, including cigarette smoking and measures of obesity, were not.

The credibility of such studies, although representative and very large, depend on the extent to which personal risk factors vary with exposure to ambient air pollution, and thus the question is whether there is confounding from omitted variables. Often these potentially confounding variables have complex interrelationships with exposure and also among the risk factors themselves. Therefore, in studies with potentially important missing covariate information, further extensions of current methods for indirect adjustment for missing variables are required to more fully characterize the dependence of exposure and health.

In this paper we propose an indirect method to adjust regression coefficients of multiple covariates accounting for multiple risk factors simultaneously that are not directly available in the primary dataset. As with previous methods, our approach assumes that there is ancillary information on important risk factors for the health endpoint, (e.g., national health surveys) that are representative of the subjects in the cohort. We examine the validity of our method by simulating a range of plausible scenarios for time to event data. As an illustration, we then apply this method to an analysis of air pollution and ischemic heart disease mortality in the Canadian census cohort study.

2. Methods

Our method of indirect adjustment is motivated by the theory of partitioned regression for linear regression models (Ruud, 2000). Let y be a vector of responses of subjects related to two sets of predictors X and U: the matrix X represents the covariates that are observed and thus available in the dataset at hand, and the matrix U represents additional covariates as confounders that are not available from the subjects in the study. We would ideally postulate a regression model of the form:

$$E\{y\}=X\beta+U\lambda, \tag{1}$$

which jointly models the two sets of covariates simultaneously and estimates two sets of unknown parameter vectors β and λ together. Our primary interest is in making inferences about some of the risk factors in X, such as air pollution, adjusting for both the other risk factors in X and U. However, we have no information on U in the current dataset and thus cannot directly calculate an unbiased estimate of β.

By the theory of partitioned regression for linear regression models we can write $\hat{\beta}$ and $\hat{\lambda}$, the least squares estimate of β and λ, respectively, as

$$\hat{\beta}=(X'X)^{-1}X'(y-U\hat{\lambda})=(X'X)^{-1}X'y-(X'X)^{-1}X'U\hat{\lambda}\equiv\hat{\gamma}-\hat{\Delta}\hat{\lambda}, \tag{2}$$

where X' is the transpose of X. The term $(X'X)^{-1}X'(y-U\hat{\lambda})$ is the least squares estimate of β based on the residual model $E\{y-U\hat{\lambda}\}=X\beta$, with $\hat{\lambda}$ from the full model in Eq. (1). We decompose this term into two further terms: $(X'X)^{-1}X'y$, which is the least squares estimate of γ defined with respect to the sub-model or reduced model $E\{y\}=X\gamma$, not including U, and $(X'X)^{-1}X'U$, which is the least squares estimate of Δ with respect to the multivariate linear model $E\{U\}=X\Delta$.

Here $\hat{\gamma}$ is the estimate of the association between the covariates available in the dataset and the response not adjusting for the set of missing covariates U, $\hat{\Delta}$ is the estimate of the multivariate relationship between the observed covariates (X) and the missing covariates (U), and $\hat{\lambda}$ is the estimate of the association between the missing covariates and the response after adjusting for the covariates in the dataset at hand.

The problem is that we cannot simultaneously estimate $\hat{\Delta}$ and $\hat{\lambda}$ from the dataset at hand and thus require ancillary information. We propose to obtain $\hat{\lambda}$ from the literature in which studies are conducted relating the risk factors U to the response y simultaneously adjusting for the risk factors X. For most cases of interest $\hat{\Delta}$ cannot be obtained from the literature. We propose to obtain $\hat{\Delta}$ from an ancillary dataset, such as national health surveys that are representative of the cohort. Of critical importance is that the amount and direction of confounding is specific to any dataset and that the amount of bias in our indirect adjustments will depend on how closely the variables in the ancillary dataset mirror both the distribution in and relationships between the variables in the dataset at hand (Breslow and Day, 1980). Thus, it is important for our method that appropriate data be found that is representative of the study population.

2.1. Indirect adjustment method for survival analysis

We focus only on cohort studies and we relate the time to event (e.g., mortality, cancer incidence) to known predictors using the Cox Proportional Hazards regression model:

$$h^{(s)}(t)=h_o^{(s)}(t)\exp\{\gamma'x\} \tag{3}$$

where $h^{(s)}(t)$ is the instantaneous probability or hazard of the occurrence of an event at time t for a subject in stratum s, γ is an unknown parameter vector relating the vector of covariates x to the hazard function with $h_o^{(s)}(t)$ the baseline hazard function defined as the hazard when $x=0$. Strata are often defined by age–sex groupings.

Although we have shown for multiple linear regression models that a simple decomposition of measured and unmeasured risk factors can be used to solve the missing data problem, the Cox model does not admit a closed-form solution. Thus, the indirect adjustment Eq. (2) can only be strictly interpreted as a partitioned regression for linear models. A partitioned regression formulation for non-linear models including the Cox model would involve partial derivatives of the log-likelihood function when forming the adjustment factors Δ. Some information contained in these derivatives, such as risk sets in a Cox partial likelihood, would not be available in an ancillary dataset. Thus Δ could not be determined explicitly. However, we argue by analogy that the above formulation for linear regression should apply. To show that in fact this analogy appears to be reasonable for many cases of interest, we carry out a series of simulations using realistic designs.

Consider that we have L covariates available in the dataset from the cohort study with the estimates of regression parameters $\hat{\gamma}$. We wish to indirectly adjust these parameter estimates for a set of R missing risk factors. Let $\tilde{U}$ be an $n\times R$ design matrix of the R risk factors for n subjects from the ancillary dataset for the missing risk factors of interest. Further let $\tilde{X}$ be an $n\times L$ design matrix of the L risk factors that are available in the cohort with values obtained from the ancillary dataset.

The indirectly adjusted parameter vector, $\tilde{\beta}$, is given by

$$\tilde{\beta}=\hat{\gamma}-(\tilde{X}'\tilde{X}^{-1})\tilde{X}'\tilde{U}\tilde{\lambda}\equiv\hat{\gamma}-\tilde{\Delta}\tilde{\lambda} \tag{4}$$

where $\tilde{\lambda}$ is a $R\times 1$ vector of the regression parameter estimates of the R risk factors on the response obtained from the literature. We note that the indirect adjustment for the lth regression parameter $\tilde{\beta}_l$ is given by $\tilde{\beta}_l=\hat{\gamma}_l-\tilde{\Delta}_{(l)}\tilde{\lambda}$, where $\tilde{\Delta}_{(l)}$ is the lth row of $\tilde{\Delta}$. Here $\tilde{\Delta}_{(l)}$ and $\tilde{\lambda}$ are independent and both random. We assume the variance of each vector component, var($\tilde{\Delta}_{(lr)}$) and var($\tilde{\lambda}_r$), is small enough to have var($\tilde{\Delta}_{(lr)}$)$*$var($\tilde{\lambda}_r$)$=0$. Then the variance of $\tilde{\beta}_l$ is given by asymptotic approximation (Goodman, 1960; Bohrnstedt and Goldberge, 1969):

$$\mathrm{var}(\tilde{\beta}_l)=\mathrm{var}(\hat{\gamma}_l)+\tilde{\Delta}_{(l)}Cov(\tilde{\lambda})\tilde{\Delta}'_{(l)}+\tilde{\lambda}'Cov(\tilde{\Delta}_{(l)})\tilde{\lambda} \tag{5}$$

with var($\hat{\gamma}_l$) obtained directly from the primary dataset analysis model. Here $Cov(\tilde{\lambda})$ is obtained from the literature and

$$Cov(\tilde{\Delta}_{(l)})=(\tilde{X}'\tilde{X})^{-1}_{(l,l)}*\tilde{\Sigma} \tag{6}$$

where

$$\tilde{\Sigma}=\tilde{U}'(I_n-\tilde{X}(\tilde{X}'\tilde{X})^{-1}\tilde{X}')\tilde{U}/n \tag{7}$$

with $(\tilde{X}'\tilde{X})^{-1}_{(l,l)}$ the lth diagonal element of $(\tilde{X}'\tilde{X})^{-1}$ and I_n an identity matrix of order$n$ (Timm, 2002). The variance of the indirectly adjusted regression parameter $\tilde{\beta}_l$ is a function of the uncertainty in the parameter estimate not adjusted based on the cohort, var($\hat{\gamma}_l$), the uncertainty in the estimates of the association between the missing risk factors and survival based on the literature, $Cov(\tilde{\lambda})$, and the uncertainty in the estimates of the association between the observed risk factors in the cohort and the missing risk factors based on ancillary dataset, $Cov(\tilde{\Delta}_{(l)})$.

For the Cox proportional hazards survival model the indirect adjustment can be written in terms of hazard ratios. Denote the hazard ratio for the lth indirectly adjusted variable by $HR_l^{adj} = \exp\{\tilde{\beta}_l\}$, the hazard ratio not adjusted for the missing covariates by$HR_l^{unadj} = \exp\{\hat{\gamma}_l\}$, and the hazard ratio of the rth missing covariate by $HR_r = \exp\{\tilde{\lambda}_r\}$. Then we have the indirectly adjusted hazard ratio

$$HR_l^{adj} = \frac{HR_l^{unadj}}{\prod_{r=1}^{R} HR_r^{\tilde{\Delta}_{(l,r)}}}, \tag{8}$$

where $\tilde{\Delta}_{(l,r)}$ is the (l,r) element of $\tilde{\Delta}$ representing the estimate of the linear association between the lth indirectly adjusted variable and the rth adjustment variable within a multivariate regression model. The amount of adjustment is dependent on the magnitude of both the hazard ratios of the adjustment variables and the association between the adjusted and adjustment variables.

2.2. Illustration 1: a dichotomous exposure variable and a dichotomous omitted variable

To further illustrate the indirect adjustment method, consider the case in which we have a dichotomous exposure, occupational exposure to a chemical for example, and want to indirectly adjust for a dichotomous variable such as current cigarette smoking. Let the hazard ratio of the exposure on some response, for example lung cancer, adjusted for age and sex but not adjusted for smoking be denoted by HR_{unadj} and the hazard ratio of current cigarette smoking on lung cancer be denoted by $HR_{smoking}$. From an ancillary dataset, suppose we know the proportion of subjects that are exposed, p_e, the proportion of subjects who smoke, p_s, and the proportion of subjects that are exposed who smoke, p_{se}. The indirect adjustment formula for this case is

$$HR_{adj} = \frac{HR_{unadj}}{HR_{smoking}^{p_{se} - p_e p_s}}. \tag{9}$$

If exposure is independent of cigarette smoking then we have $p_{se} = p_e p_s$ and the adjusted and unadjusted hazard ratios are the same. If proportionally more subjects in the exposed group are cigarette smokers compared to the unexposed group, then $p_{se} > p_e p_s$ and the effect of the indirect adjustment would be to reduce the hazard ratio. Similarly, if proportionally fewer subjects in the exposed group were cigarette smokers then $p_{se} < p_e p_s$ and the adjusted hazard would be larger than the unadjusted hazard ratio.

2.3. Illustration 2: a continuous exposure variable and a continuous omitted variable

Now consider the case of a single continuous variable, x, whose regression coefficient is to be adjusted and a single continuous adjustment variable, u. Then the indirect adjustment formula is given by

$$\tilde{\beta} = \hat{\gamma} - \hat{\rho}\left(\frac{s_u}{s_x}\right)\tilde{\lambda} \tag{10}$$

where $\hat{\rho}$ is the empirical Pearson correlation between x and u with s_u and s_x the standard deviations of u and x respectively (Montgomery et al., 2006). When $s_u = s_x$, the indirect adjustment formula written in terms of hazard ratios is

$$HR^{adj} = \frac{HR^{unadj}}{(HR_u)^{\hat{\rho}}} \tag{11}$$

where HR_u is the hazard ratio for the adjustment variable u. If x and u are uncorrelated (i.e. $\hat{\rho} = 0$), then $HR^{adj} = HR^{unadj}$. If x and u are positively correlated, then $HR^{adj} < HR^{unadj}$ and if negatively correlated then $HR^{adj} > HR^{unadj}$.

3. Results

3.1. Simulation study

We assessed the validity of our indirect adjustment method using a simulation study whose details are given in Appendix. Briefly, we considered two variables x (i.e., air pollution) and u (i.e., smoking), with x as the adjusted variable and u as the adjustment variable. One ten thousand realizations of these two variables were generated assuming a standard bivariate normal distribution with correlation either 0.2 or 0.5. For each pair of (x, u) we simulated 30,000 event times from a Weibull distribution with scale parameter defined as a log-linear function of x and u. We varied the shape parameter of the Weibull distribution such that the baseline hazard function increased with the power of follow-up time. We selected values of the shape parameter from 1 to 5, where the power of time is the shape value minus 1. For example, when the shape parameter equals unity the baseline hazard is a constant, as the hazard ratio is no longer dependent on time. We also simulated Weibull censoring times with 0.9 and 0.5 censoring rates such that approximately 10 or 50% of the subjects experienced an event. The unknown regression parameter associated with x was defined such that the hazard ratio of the parameter multiplied by the negative of the shape parameter evaluated at the range of x was 1.5, a value typical of hazard ratios in mortality studies of air pollution. The unknown parameter associated with u was defined such that the ratio of the hazard ratio relating x to a response and the hazard ratio relating u to the response were 1,2,4,8, or 16. For example, if the hazard ratio for air pollution was 1.5, the hazard ratio for current cigarette smoking could be as large as 1.5 × 16=24.

To obtain estimates, we applied Weibull regression model and Cox Proportional-Hazards (Cox PH) model to both the full and reduced models. The risk estimates from both models were quite close to each other, and thus we report the risk estimates from Cox PH model only.

We summarize the adequacy of our method by calculating relative bias of adjusted risk $(|\tilde{\beta} - \hat{\beta}|/\beta)$, and unadjusted risk $(|\hat{\gamma} - \hat{\beta}|/\beta)$, where $\tilde{\beta}$ and $\hat{\gamma}$ are the respective adjusted and unadjusted estimates, $\hat{\beta}$ is the estimate from full model, which is believed as the best estimate, and β is the true value set up for the simulation. The bias represents the mean difference among the 30,000 simulations of the estimate of the parameter associated with x in the full model including u, and the corresponding indirectly adjusted parameter estimate, divided by the true value of the parameter. These relative biases are summarized in Table 1 by the ratio of the hazard ratios between x and u (1,2,4,8, or 16), the censoring rate (0.9 or 0.5), and the correlation between x and u (0.2 or 0.5).

The relative bias was insensitive to the Weibull shape parameter (see Table A1), but increased as the ratio of the hazards ratios of the two variables, the correlation among variables, and the censoring rate increased. Table 1 summarizes the amount of bias by the censoring rate, hazards ratios, and correlation as averaged over all shape parameter values. As expected, the reduced model mostly over-estimated the risk, but the adjusted risk estimates were close to the full model risk estimates. The relative bias in the unadjusted HRs of x is also reported in

Table 1
Percent relative bias (difference between adjusted or unadjusted risk estimate and full model risk estimate compared to the the true risk value) by censoring rate and hazard ratio of two regression coeeficients for two correlations between the variables (cor=0.2 and cor=0.5).

Censoring rate	Hazard ratio	Adjusted[a]		Unadjusted[b]	
		cor=0.2	cor=0.5	cor=0.2	cor=0.5
0.5	1	0.1	0.1	20.1	45.2
	2	1.2	1.4	53.5	121.3
	4	3.6	4.8	85.5	195.3
	8	8.0	11.1	115.7	266.4
	16	14.5	20.9	143.6	333.9
0.9	1	0.3	0.2	19.9	45.1
	2	2.3	2.6	52.4	120.0
	4	7.0	8.8	82.1	191.1
	8	14.8	20.0	108.8	257.4
	16	25.6	36.6	132.4	318.4

[a] $(|\tilde{\beta} - \hat{\beta}|/\beta) \times 100$, where $\tilde{\beta}$ is the adjusted estimates, $\hat{\beta}$ is the estimate from full model, which is believed as the best estimate, and β is the true value.

[b] $(|\hat{\gamma} - \hat{\beta}|/\beta) \times 100$, where $\hat{\gamma}$ is the unadjusted estimates from the reduced model, $\hat{\beta}$ is the estimate from full model, which is believed as the best estimate, and β is the true value.

Table 1. These relative biases are much larger than their adjusted counterparts demonstrating the effect of the indirect adjustment approach.

3.2. An example of fine particulate air pollution and mortality in a national cohort study

The association between long-term exposure to ambient concentrations of fine particulate matter (particles with aerodynamic diameter less than 2.5 μm) and cause-specific mortality has been estimated in Canada in a subset of the Census Cohort (Crouse et al., 2012). The cohort was composed of Canadians 25 years of age and older who completed the 1991 Census long form (20% of population) and whose records were subsequently linked to the Canadian Mortality Database (from June 4, 1991 to December 31, 2001) using deterministic and probabilistic linkage methods (Wilkins et al., 2008). In this example, we included only those subjects who were non-immigrants, leaving approximately 2.1 million subjects. Immigrants to Canada are healthier and thus survive longer than native born Canadians (Wilkins et al., 2008). In addition, they tend to live in larger cities with higher pollution exposures (Crouse et al., 2012). We assigned 2001–2006 average concentrations of fine particulate matter to each subject's home address six-character postal code in 1991 based on satellite remote sensing observations (van Donkelaar et al., 2010). The six-character postal code represents a block face in cities but can represent a much larger area in rural settings.

Several mortality risk factors recorded on the long-form census were included in the survival model (i.e. income, education, occupation, marital status, aboriginal status, employment status, visible minority, and size of community). The baseline hazard function was stratified by single year age groups and sex. However, cigarette smoking habits and obesity status, two important risk factors for ischemic heart disease mortality, were not available.

We wished to indirectly adjust the regression coefficient for fine particulate matter for these two missing covariates by characterizing cigarette smoking habits using two binary variables: former versus never cigarette smoker and current versus never smoker. As well, body mass index (kg/m^2) was characterized using four binary variables describing ranges 25–30, 30–35, 35–40, and >40 kg/m^2 compared to <25 kg/m^2. We obtained from the American Cancer Society Cancer Prevention II (ACS) cohort (Pope et al., 2004) hazard ratio estimates for current versus never smokers (HR=2.03; 95%CI: 1.96–2.10) and former smokers (HR=1.35; 95%CI: 1.29–1.37). We also obtained an estimate of the hazard ratio of mortality due to ischemic heart disease associated with body mass index (Prospective Studies Collaboration, 2009). The hazard ratio per 5 kg/m^2 increase in body mass index above 25 kg/m^2 was 1.39 (95%CI: 1.34–1.44). We then calculated the hazard ratio based on the difference between the group mean body mass index from our ancillary dataset (see below) and 25 kg/m^2 (Table 2).

The association between the variables that were included in the survival model (age, sex, fine particulates, income, education, occupation, marital status, aboriginal status, employment status, visible minority, and size of community) and the six indirect adjustment variables was also required. This relationship was estimated using the Canadian Community Health Survey (Statistics Canada, 2003), a bi-annual, national, population-based cross-sectional survey of Canadians that started in 2001. We first assigned the remote sensing-based concentrations of fine particulate matter to the centroid of the home address of the six-character postal code of all subjects in the 2001, 2003, and 2005 panels (sample size of 188,617 subjects) of the Canadian Community Health Survey who were 25 years of age or older and who were born in Canada. These panels were selected to coincide with the 2001–2006 average fine particulate matter concentrations.

We included in the design matrix, $\tilde{X}$, data from the Canadian Community Health Survey for the same variables and category definitions as in the survival model applied to the census cohort. We added a column of 1s to represent the baseline hazard function and indicator variables for age–sex interactions to represent the stratification of the baseline hazard by age and sex.

The elements of the $\tilde{\Delta}$ matrix corresponding to fine particulate matter are presented in Table 2 for three scenarios. In the first scenario we included a column of 1s and fine particulate matter concentrations only, denoted by ***None***. In the second scenario we also included indictor variables for the age–sex interaction, denoted by ***Age–Sex***. In the third scenario we additionally included all the variables that were included in the survival model, denoted by ***All Variables***.

Negative associations were observed between concentrations of fine particulate matter and both current and former cigarette smokers for the ***None*** and ***Age–Sex*** scenarios (Table 2). However, the association decreased by an order of magnitude for the ***All Variables*** scenario. We observed a negative association between concentrations of fine particulate matter and all four body mass index categories for the ***None*** scenario. However, these associations were null for the two lowest body mass index categories for both the ***Age–Sex*** and ***All Variables*** scenarios (Table 2). The association between fine particulate matter and the two highest body mass index categories decreased by several orders of magnitude for both the ***Age–Sex*** and ***All Variables*** scenarios compared to the ***None*** scenario. Including all the variables in the indirect adjustment that were included in the survival model appears to have explained most of the association between fine particulate matter and all six adjustment variables.

The indirectly adjusted hazard ratio for an increase of 10 μg/m^3 in fine particulate matter was substantially larger (HR=1.82; 95% CI: 1.73–1.90; for the ***None*** scenario compared to the hazard ratio without any indirect adjustment (HR=1.31; 95% CI: 1.27–1.34). This was due to the strong negative associations between fine particulate matter and either cigarette smoking or body mass index (Table 2). The indirectly adjusted hazard ratio under the ***Age–Sex*** scenario was smaller (HR=1.36; 95% CI: 1.32–1.41) compared to the ***None*** scenario, mostly due to the much weaker association between fine particulate matter and all four categories of body mass index. The indirect adjustment had little effect on the hazard ratio (HR=1.32; 95% CI: 1.28–1.36) for the ***All Variables*** scenario compared to the hazard ratio without any indirect adjustment, since these additional variables were explaining much of the association between fine particulate matter and the adjustment variables.

The standard error of the indirectly adjusted regression coefficient, $\tilde{\beta}$, increased by 69%, 7%, and 2% for the ***None***, ***Age–Sex***, and ***All Variables*** scenarios respectively compared to the standard error of the coefficient not indirectly adjusted, $\hat{\gamma}$. Reductions in both the adjustment values, $\tilde{\Delta}_{(I)}$, and uncertainty in these values, resulted in smaller standard errors as the number of variables contained in the adjustment matrix, $\tilde{X}$, increased.

4. Discussion

We proposed a new methodology based on partitioned regression to indirectly adjust risk estimates for potentially important confounding variables that are missing. Our methods incorporate indirect adjustment for several missing confounding variables simultaneously in addition to controlling for the relationship between observed variables of primary interest and missing variables. We placed no restrictions on the form of the primary variables (continuous, categorical) or the form of the missing

Table 2
Quantities for smoking and body mass index (BMI) required for indirectly adjusting the association between mortality from ischemic heart disease and concentrations of fine particulate matter.

Missing risk factor	Percent in the Canadian community health survey	Log-hazard ratio (standard error)	Associations between smoking and BMI with concentrations of fine particulate matter from the Canadian community health survey		
			Variables Included in Adjustment Model		
			None	***Age–Sex***	***All variables***
Smoker reference category					
Never smoker	25.7	NA	NA	NA	NA
Current smoker	30.5	0.70804 (0.00031)	−0.003645	−0.004032	0.000746
Former smoker	43.8	0.30010 (0.00024)	−0.006530	−0.005597	−0.000746
BMI reference category					
BMI $<$ 25 kg/m^2	44.6	NA	NA	NA	NA
25 $\leq$ BMI $<$ 30 (27.25 kg/m^2)[a]	35.9	0.14842 (0.00008)	−0.008664	~0	~0
30 $\leq$ BMI $<$ 35 (31.97 kg/m^2)[a]	13.7	0.45742 (0.00039)	−0.009512	~0	~0
35 $\leq$ BMI $<$ 40 (36.93 kg/m^2)[a]	4.0s	0.78390 (0.00195)	−0.010633	0.000234	−0.000392
BMI $\geq$ 40 (44.52 kg/m^2)[a]	1.8	1.28647 (0.00515)	−0.011261	0.000005	−0.000644

[a] BMI group mean.

variables (continuous, categorical). We obtained closed form expressions for the variance of the adjusted parameter estimates, Eq. (5), thus alleviating the need to use simulation approaches as suggested previously (Steenland and Greenland, 2004).

Based on the results of our simulation study, indirect adjustment approach yielded only a small amount of relative bias less than 20% for all realistic scenarios examined. For each death time, the covariates of the subject who experienced the fatality are compared to the covariates of the set of subjects alive at that time for the Cox partial likelihood. Thus only a small subset of covariate information is assigned to subjects who die when the censoring rate is very high, such as 0.9. However, the covariate values of all subjects are included in the indirect adjustment formula based on information obtained by the ancillary dataset. Even if the covariate information obtained by the ancillary data are in fact representative of the corresponding covariate information from the entire cohort, that subset of information based on those subjects who died may not be as representative.

We also note that the correlation between fine particulate matter concentrations and the six indicator variables representing cigarette smoking habits or BMI that were used in the indirect adjustment for the Canadian Census cohort ranged from −0.04 to −0.02. We would then expect little bias in our indirect adjustments based on these very modest correlations in the example presented.

An important aspect of this approach depends on the representativeness of the ancillary information. Representativeness, like validity, is based on whether the population that provided the ancillary data is drawn from the same target population as the cohort or is otherwise similar in important respects, such as age, sex, health status, and geographic coverage. As well, similarity between studies in the type of data that has been collected will also be important; e.g., similar questions on income, education, occupation. In our example of air pollution and mortality, we were able to select subjects in the ancillary dataset with similar characteristics (e.g., age, immigration status, geographic areas), to assign to each subject in the ancillary dataset concentrations of air pollution that were based on the same exposure model as we used in the primary dataset, and the definitions of the available covariates were the same in the Census Cohort and the Canadian Community Health Survey.

Our indirect adjustment method estimates the association between the missing factors and the available factors contained in the survival model using ancillary information. We do not attempt to estimate the missing risk factors directly from the ancillary information as has been suggested (Mason et al., 2012). The accuracy of the missing risk factor estimate model is dependent on the quality of information needed to predict the missing risk factor, which may be limited for the dataset at hand. For example, it is likely that poor predictions of missing risk factors would be obtained if covariate information from the dataset at hand is limited to a few variables such as age and sex. Since our method does not rely on predicting missing information we are not subject to this limitation.

We make the following recommendations on assessing the adequacy of the ancillary health studies in representing the cohort study. We first suggest that the distribution of air pollution among subjects in both the cohort and health survey be examined. Second, the correlation amongst the variables available in the cohort (i.e. education, income) should be examined and compared to the correlation amongst the same variables in the health survey. Concerns should be raised about using the ancillary data if these distributions or correlations are not similar. Third, survival models could be examined for which specific variables are excluded. The corresponding air pollution Hazard Ratio could be indirectly adjusted for those excluded variables using the ancillary health survey and compared to the Hazard Ratio based on a survival model consisting of both the variables included in the survival model and those excluded. The Hazard Ratio of the excluded covariates, needed for the indirect adjustment, could be obtained from the survival model with complete representation of the all the covariates. The air pollution Hazard ratio estimate based on the full model should be similar to that based on the reduced survival model after indirect adjustment.

We suggest that our indirect adjustment approach could be applied to cases other than a survival model. For example for logistic or Poisson regression models as long as the covariate information enters the model as a linear combination as is the case here. Clearly, this hypothesis needs to be supported by appropriate simulation studies.

In summary, we proposed a new method to indirectly adjust risk estimates obtained from survival models for multiple missing covariates (either continuous or categorical) simultaneously. We have demonstrated by simulation that this method performs adequately in correcting bias from missing covariates in most situations of interest in environment epidemiology.

Information on funding sources

This research article does not have any funding source and did not involve either humans or animals.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.envres.2014.05.016.

References

Axelson, O., 1980. Aspects of confounding and effect modification in the assessment of occupational cancer risk. J. Toxicol. Environ. Health 6, 1127–1131.

Breslow, N.E., Day, N.E., 1980. Statistical Methods in Cancer Research. Volume I – The Analysis of Case-control Studies. IARC Scientific Publications. No. 32, Lyon.

Bohrnstedt, G.W., Goldberge, A.S., 1969. On the exact covariance of products of random variables. J. Am. Stat. Assoc. 64 (328), 1439–1442.

Crouse, D.L., Peters, P.A., van Donkelaar, A., et al., 2012. Risk of mortality in relation to long-term exposure to low concentrations of fine particulate matter: a Canadian national-level cohort study. Environ. Health Perspect. 120, 708–714.

Goodman, L.A., 1960. On the exact variance of products. J. Am. Stat. Assoc. 55, 708–713.

Langholz, B., Goldstein, L., 1996. Risk set sampling in epidemiologic cohort studies. Stat. Sci. 11 (1), 35–53.

Liddel, F.D.K., McDonald, J.C., Thomas, D.C., 1977. Methods for cohort analysis: appraisal byapplication to asbestos mining (with discussion). J. R. Stat. Soc. A 140, 469–490.

Mason, A., Richardson, S., Plewis, I., Best, N., 2012. Strategy for modelling nonrandom missing data mechanisms in observational studies using Bayesian methods. J. Off. Stat. 28, 279–302.

Montgomery, D.C., Peck, E.A., Vining, G.G., 2006. Introduction to Linear Regression Analysis, 4th ed John Wiley & Sons Inc, New Jersey, pp. 49–50.

Pope 3rd, C.A., Burnett, R.T., Thurston, G.D., et al., 2004. Cardiovascular mortality and long-term exposure to particulate air pollution: epidemiological evidence of general pathophysiological pathways of disease. Circulation 109 (1), 71–77.

Prospective Studies Collaboration, 2009. Body-mass index and cause-specific mortality in 900,000 adults: collaborative analyses of 57 prospective studies. Lancet 373 (9669), 1083–1096.

Ruud, P.A., 2000. An Introduction to Classical Econometric Theory. Oxford University Press, New York, USA.

Statistics Canada, 2003. Canadian Community Health Survey: User Guide for the Public Use Microdata File. Ottawa, Health Statistics Division, Statistics Canada.

Steenland, K., Greenland, S., 2004. Monte Carlo sensitivity analysis and Bayesian analysis ofsmoking as an unmeasured confounder in a study of silica and lung cancer. Am. J. Epidemiol. 160 (4), 384–392.

Timm, N.H., 2002. Applied Multivariate Analysis. Springer, New York, pp. 106–115.

Villeneuve, P.J., Goldberg, M.S., Burnett, R.T., et al., 2010. Associations between cigarette smoking, obesity, sociodemographic characteristics, and remote sensing derived estimates of ambient PM2.5: results from a Canadian population-based survey. Occup. Environ. Med. 68 (12), 920–927.

Wilkins, R., Tjepkema, M., Mustard, C., et al., 2008. The Canadian census mortality follow-up study, 1991 through 2001. Health Rep. 19 (3), 25–43.

van Donkelaar, A., Martin, R.V., Brauer, M., et al., 2010. Global estimates of ambient fine particulate matter concentrations from satellite-based aerosol optical depth: development and application. Environ. Health Perspect. 118 (6), 847–855.