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
0013-9351/

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 $\tilde{\lambda}$ 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 $\tilde{\Delta}$ 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, $\mathrm{var}(\tilde{\Delta}_{(lr)})$ and $\mathrm{var}(\tilde{\lambda}_r)$, is small enough to have $\mathrm{var}(\tilde{\Delta}_{(lr)}) * \mathrm{var}(\tilde{\lambda}_r) = 0$. 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 $\mathrm{var}(\hat{\gamma}_l)$ 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, $\mathrm{var}(\hat{\gamma}_l)$, 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