Available online at www.sciencedirect.com
ScienceDirect
Journal of Statistical Planning and Inference 137 (2007) 449–461
journal of statistical planning and inference
www.elsevier.com/locate/jspi

Regression calibration for logistic regression with multiple surrogates for one exposure

Edie A. Weller[a,*], Donald K. Milton[b], Ellen A. Eisen[b], Donna Spiegelman[c]

[a] *Department of Biostatistics, Harvard School of Public Health and Dana Farber Cancer Institute, 44 Binney Street Mayer 225, Boston, MA 02115, USA*

[b] *Department of Environmental Health, Harvard School of Public Health, 665 Huntington Avenue, Boston, MA 02115, USA*

[c] *Department of Biostatistics and Epidemiology, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115, USA*

Received 24 June 2005; received in revised form 25 January 2006; accepted 26 January 2006
Available online 30 March 2006

Abstract

Methods have been developed by several authors to address the problem of bias in regression coefficients due to errors in exposure measurement. These approaches typically assume that there is one surrogate for each exposure. Occupational exposures are quite complex and are often described by characteristics of the workplace and the amount of time that one has worked in a particular area. In this setting, there are several surrogates which are used to define an individual's exposure. To analyze this type of data, regression calibration methodology is extended to adjust the estimates of exposure-response associations for the bias and additional uncertainty due to exposure measurement error from multiple surrogates. The health outcome is assumed to be binary and related to the quantitative measure of exposure by a logistic link function. The model for the conditional mean of the quantitative exposure measurement in relation to job characteristics is assumed to be linear. This approach is applied to a cross-sectional epidemiologic study of lung function in relation to metal working fluid exposure and the corresponding exposure assessment study with quantitative measurements from personal monitors. A simulation study investigates the performance of the proposed estimator for various values of the baseline prevalence of disease, exposure effect and measurement error variance. The efficiency of the proposed estimator relative to the one proposed by Carroll et al. [1995. Measurement Error in Nonlinear Models. Chapman & Hall, New York] is evaluated numerically for the motivating example. User-friendly and fully documented Splus and SAS routines implementing these methods are available (`http://www.hsph.harvard.edu/faculty/spiegelman/multsurr.html`).

Keywords: Measurement error; Validation study; Main study; Metal working fluids; Exposure

1. Introduction

Many researchers have proposed methods to adjust for errors in exposure measurement (Rosner et al., 1989, 1990, 1992; Armstrong, 1990; Carroll and Stefanski, 1990; Thomas et al., 1993; Kuha, 1994; Carroll et al., 1995; Lee and Sepanski, 1995; Lyles and Kupper, 1997). Two related methods have been referred in the literature to as regression calibration. The first approach was proposed by Rosner et al. (1989, 1990) and involves using a surrogate for the true

* Corresponding author. Tel.: +1 617 632 2445; fax: +1 617 632 2444.
E-mail address: eweller@jimmy.harvard.edu (E.A. Weller).

0378-3758/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2006.01.009

exposure when fitting the primary regression model, i.e., the one which includes the parameter of interest, β. Based on asymptotic theory, the estimated parameters from this surrogate model were shown to converge, approximately, to functions of the parameters of interest, upon which corrections were derived. A second approach was proposed by Carroll and Stefanski (1990). This approach involves substituting the conditional expectation of the true exposure given the observed surrogate and other covariates into the primary regression model, and using bootstrap or sandwich methods to adjust the standard errors of the estimates. Thurston et al. (2003) showed that these two methods are identical under fairly general circumstances, for main study/external validation study designs. In such designs, the main study includes data on the outcome of interest, surrogates of the exposure and confounding covariates. The external validation study includes data about the exposure, the surrogates and confounding covariates, but does not contain outcome data. The external validation studies may be conducted independent of the main study as long as it can be assumed that the measurement error model parameters to be used in subsequent measurement error correction of results from the main study but estimated in the validation study are "transportable" to the main study. The approach proposed in this paper follows that of Rosner et al. (1989, 1990).

In Rosner et al.'s regression calibration for main study/external validation study designs, the point and interval estimates of association are first obtained by fitting a logistic regression model

$$\text{logit}[\Pr(D_i = 1)] = \alpha_0 + \mathbf{W_i}\boldsymbol{\alpha}_1 + \mathbf{Z_i}\boldsymbol{\alpha}_2, \tag{1}$$

where $\mathbf{W_i}$ is a vector of r surrogates for exposure X_i for individual i ($i = 1, 2, \ldots, n_1$) in the main study, $\mathbf{Z_i}$ is a vector of the s covariates measured without error and the vectors $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ represent uncorrected log odds ratios describing a one unit increase in the model covariates. The parameter vectors $\boldsymbol{\alpha}_1' = (\alpha_{11}, \alpha_{12}, \ldots, \alpha_{1r})$ and $\boldsymbol{\alpha}_2' = (\alpha_{21}, \alpha_{22}, \ldots, \alpha_{2s})$ correspond to the covariates measured with and without error, respectively.

The logistic regression coefficients from Eq. (1) can then be adjusted for bias due to measurement error in a one-step procedure to obtain estimates of $\boldsymbol{\beta}$ from the model $\text{logit}[\Pr(D_i = 1)] = \beta_0 + \mathbf{X_i}\boldsymbol{\beta}_1 + Z_i\boldsymbol{\beta}_2$. Rosner et al. (1989, 1990) proposed that the point and interval estimates of log odds ratio can be corrected for measurement error using the formula $\hat{\boldsymbol{\beta}}_{\text{RC}} = \hat{\Gamma}_{\text{RC}}^{-1}\hat{\boldsymbol{\alpha}}$ where the $(r + s)$ vector $\hat{\boldsymbol{\alpha}} = (\hat{\boldsymbol{\alpha}}_1, \hat{\boldsymbol{\alpha}}_2)$ is estimated from (1) in the main study and $\hat{\boldsymbol{\beta}}_{\text{RC}}$ are the corrected logistic regression coefficients. The $(r + s) \times (r + s)$ matrix Γ_{RC} is estimated in the validation study by fitting the linear regression model

$$\mathbf{X_i} = \gamma_0 + \mathbf{W_i}\boldsymbol{\gamma}_1 + \mathbf{Z_i}\boldsymbol{\gamma}_2 + \varepsilon_i, \tag{2}$$

where $\mathbf{X_i}$ are the perfectly measured exposure variables for individual i

$$(i = 1, 2, \ldots, n_2), \quad \Gamma_{\text{RC}} = \begin{pmatrix} \gamma_1 & \mathbf{0} \\ \gamma_2 & \mathbf{I} \end{pmatrix}, \quad \boldsymbol{\gamma}_1' = (\gamma_{11}, \gamma_{12}, \ldots, \gamma_{1r}), \quad \boldsymbol{\gamma}_2' = (\gamma_{21}, \gamma_{22}, \ldots, \gamma_{2s}),$$

and $\boldsymbol{\epsilon}_i$ is a random error with mean 0 and variance $\sigma^2_{X|\mathbf{W},\mathbf{Z}}$. The variance of $\hat{\boldsymbol{\beta}}_{\text{RC}}$ was derived using the multivariate delta method. As shown by Kuha (1994), this method is valid if either the outcome of interest is rare and $\boldsymbol{\epsilon}_i$ is Gaussian, or the measurement error is not severe, that is, if $\boldsymbol{\beta}_1'\Sigma_{X|\mathbf{W},\mathbf{Z}}\boldsymbol{\beta}_1$ is small. In addition, this method assumes that for each exposure ($\mathbf{X}$), a single surrogate ($\mathbf{W}$) is measured, that is, that the dimension of $\mathbf{X}$ and $\mathbf{W}$ are equal. In internal validation studies with outcome measured as well as the surrogates, exposure and confounding covariates, $\hat{\boldsymbol{\beta}}_{\text{RC}}$ can be combined with the maximum likelihood estimator obtained from the internal validation study using an inverse-variance weighted summary estimator, as long as sampling into the validation study is conditionally independent of X given ($\mathbf{W}$, $\mathbf{Z}$) (Spiegelman et al., 2001).

In many occupational studies, multiple factors serve as surrogates for a single exposure. Exposure is generally described by characteristics of the workplace and the amount of time one has worked in particular areas. Often an industrial hygenist will conduct a detailed exposure assessment using personal or area monitors, resulting in a more quantitative measure of exposure. Without these more accurate exposure measurements, the use of alternative exposure assessments can result in substantial misclassification, reducing or obscuring the exposure–effect relationship (Smith, 1987). In small studies of acute health outcomes, current exposure may be measured for each subject. More commonly, personal exposure is measured only on a subset of the subjects and these values are then used to estimate average exposure by job or exposure zone. In retrospective studies of chronic diseases, long-term exposures are generally estimated by a weighted sum of many of these averages. Typically, individual exposure levels are assessed by classifying workers

as exposed/not exposed or by assigning each worker an average exposure for their particular work area. The average values are used directly in the model relating exposure to the health outcome without any adjustment for the fact that the exposure values are estimated. Health effects models are also fit using the job characteristics as surrogates for exposure.

To accommodate the structure of these type of studies, a regression calibration approach is proposed which extends the methods of Rosner et al. (1989,1990). The quantitative exposure measure is assumed to be related to the health outcome by a logistic model. The regression of the quantitative exposure on the characteristics of the workplace is assumed to be linear. Results of a simulation study to assess the properties of the proposed estimator with small sample size and under deviations from the rare disease assumption are presented. The asymptotic efficiency of the proposed estimator relative to the corresponding Carroll estimator is evaluated numerically in a broad region of the parameter space centered around the parameters of the motivating example. Finally, we apply the proposed approach to an epidemiologic study assessing the relationship between the exposure to metal working fluids and respiratory function in the presence of measurement error.

2. A motivating example

An epidemiologic study jointly sponsored by the United Automobile Workers Union and the General Motors Corporation evaluated the relationship between current exposure to metal working fluids (MWF) and respiratory function (Greaves et al., 1997). This study was part of a larger evaluation that included an assessment of past MWF exposure and cancer related mortality (Eisen et al., 1992; Hallock et al., 1994; Tolbert et al., 1992; Woskie et al., 1994). In this study, automobile workers from three General Motors facilities, who were currently exposed to one of three primary types of MWF (straight MWF, soluble MWF and synthetic MWF) were recruited along with assembly workers, who were treated as the low-exposed group because they had no direct exposure to MWF, for a total of 1811 workers included. Information on exposure included job characteristics such as plant, the type of metal-working machine (grinding and non-grinding) and the metal working fluid type. In an exposure assessment study by Woskie et al. (1994), full shift (8 h) personal samples of aerosol exposure in the breathing zone of 475 workers were collected. The exposure based on the thoracic aerosol fraction (i.e., the sum of the two smallest size fractions measured with the personal monitors) in mg/m^3 measures the intensity of exposure to MWF aerosol. Exposure variables for each type of fluid was defined for all 1811 individuals in four different ways: (1) exposed and low-exposed, (2) low-exposed and tertiles of exposure, (3) arithmetic mean of the tertile exposure category and (4) individual estimate based on job characteristics and area aerosol concentration (Greaves et al., 1997). The aim of this paper is to obtain a single estimate of the exposure effect on the respiratory function which combines information from each of the job characteristics and accounts for the variability associated with the estimated exposure variables.

3. Methods

Suppose that the true exposure (X) and the perfectly measured covariates ($\mathbf{Z}$) are related to the probability of binary outcome (D) by the logistic function $\text{logit}[\Pr(D=1)]=\beta_0+X\beta_1+\mathbf{Z}\boldsymbol{\beta}_2$ where $\boldsymbol{\beta}_2'=(\beta_{21}, \beta_{22}, \ldots, \beta_{2s})$. In addition, we assume that the linear regression model given in (2) is appropriate to relate the r surrogates of $\mathbf{W}$ and the s covariates of $\mathbf{Z}$ to the true exposure. Standard regression diagnostic methods can be used to check the validity of the linear regression model (Belsey et al., 1980). The selection of surrogate and confounding covariates into the models should involve discussions with collaborating investigators, who provide valuable insight regarding the importance of the variables in the area of application, regardless of the statistical significance of the variables in the specific data set at hand. The goal is to obtain point and interval estimates of $\boldsymbol{\beta}$ and $e^{\boldsymbol{\beta}}$, which relate exposure (X) to outcome (D), adjusting for the covariates ($\mathbf{Z}$). The problem is that the true exposure, or an unbiased estimate of the true exposure, is not measured on all subjects. Instead, we have multiple surrogates for exposure, denoted $\mathbf{W}$, that are measured on n_1 subjects in the main study, and X and $\mathbf{W}$ measured on n_2 additional validation study subjects.

To estimate the exposure effects, i.e., to estimate the regression coefficients, $\boldsymbol{\beta}'=(\beta_0, \beta_1, \boldsymbol{\beta}_2')$, we propose the following regression calibration approach, which requires the availability of external validation data. External validation study includes surrogate and exposure data as well as confounding covariates. The derivation of this estimator follows that of Rosner et al. (1989).

We assume that the measurement error model given by (2) is appropriate and that the sample disease probability is small, i.e.

$$\Pr(D|\mathbf{W}, \mathbf{Z}) \approx \int_x \mathrm{e}^{\beta_0 + X\beta_1 + \mathbf{Z}\boldsymbol{\beta}_2} f_{X|\mathbf{W},\mathbf{Z}}(X|\mathbf{W}, \mathbf{Z})\, \mathrm{d}x,$$

where $f_{X|\mathbf{W},\mathbf{Z}}$ is the normal density function of X given $\mathbf{W}, \mathbf{Z}$ with mean $\mu_{X|\mathbf{W},\mathbf{Z}}$ and variance $\sigma^2_{X|\mathbf{W},\mathbf{Z}}$. After some simple algebra and completing the square,

$$\Pr(D|\mathbf{W}, \mathbf{Z}) \approx \mathrm{e}^{\beta_0 + \frac{1}{2}\beta_1^2\sigma^2_{X|\mathbf{W},\mathbf{Z}}/2 + \mu_{X|\mathbf{W},\mathbf{Z}}\beta_1 + \mathbf{Z}\boldsymbol{\beta}_2}.$$

If the measurement error model (2) fits the data, that is if

$$\mu_{X|\mathbf{W},\mathbf{Z}} = \gamma_0 + \mathbf{W}\boldsymbol{\gamma}_1 + \mathbf{Z}\boldsymbol{\gamma}_2, \tag{3}$$

then the probability of the disease outcome given $\mathbf{W}$ and $\mathbf{Z}$ is given by $\Pr(D|\mathbf{W}, \mathbf{Z}) \approx \mathrm{e}^{\alpha_0 + \mathbf{W}\boldsymbol{\alpha}_1 + \mathbf{Z}\boldsymbol{\alpha}_2}$ where $\alpha_0 = \beta_0 + \frac{1}{2}\beta_1^2\sigma^2_{X|\mathbf{W},\mathbf{Z}}/2 + \gamma_0\beta_1$, $\boldsymbol{\alpha}_1 = \beta_1\boldsymbol{\gamma}_1$ and $\boldsymbol{\alpha}_2 = \beta_1\boldsymbol{\gamma}_2 + \boldsymbol{\beta}_2$. Using the estimates of $\boldsymbol{\alpha}' = (\alpha_0, \boldsymbol{\alpha}_1', \boldsymbol{\alpha}_2')$ obtained from the model of the probability of disease outcome on $\mathbf{W}$ and $\mathbf{Z}$ and estimates of $\boldsymbol{\gamma}' = (\gamma_0, \boldsymbol{\gamma}_1', \boldsymbol{\gamma}_2')$ obtained from the measurement error model of X on $\mathbf{W}$ and $\mathbf{Z}$, an estimated exposure effect for each surrogate is determined by solving the equations $\boldsymbol{\alpha}_1 = \boldsymbol{\beta}_1\boldsymbol{\gamma}_1$ for $\boldsymbol{\beta}_1$, giving $\beta_{\mathbf{X}}$.

This vector of correlated, approximately consistent estimates of $\beta_1, \beta_{\mathbf{X}}$, are combined using generalized least squares theory to obtain the estimated exposure effect parameter $\hat{\beta}_1$ which has minimum variance among all unbiased linear combinations of the individual estimates, as the inverse estimated variance weights approach their true values, i.e. as long as they are accurately estimated (Arnold, 1981, p. 201).

This final summary estimated exposure effect has the desirable feature of being in the units of X. In addition, the adjusted parameters for the intercept $\hat{\beta}_0$ and the perfectly measured covariates $\hat{\boldsymbol{\beta}}_2$ are estimated. This approach assumes normality of ε and rare sample disease prevalence. Similar results are shown to hold assuming a small $\beta_1^2\sigma^2_{X|\mathbf{W},\mathbf{Z}}$ (that is, small measurement error and/or small effect of X on D) using a second order Taylor series expansion (Appendix A). The quantity $\beta_1^2\sigma^2_{X|\mathbf{W},\mathbf{Z}}$ is referred to as the parameter of the small measurement error approximation.

The following details the steps for implementing this approach:

1. Fit the logistic regression model given in (1) of D on surrogates $\mathbf{W}$ and perfectly measured covariates $\mathbf{Z}$ in the n_1 subjects in the main study to obtain estimates of $\boldsymbol{\alpha}' = (\alpha_0, \boldsymbol{\alpha}_1', \boldsymbol{\alpha}_2')$.
2. Fit the measurement error model given in (2) among the n_2 external validation study subjects using ordinary least square regression to find estimates of $\boldsymbol{\gamma}' = (\gamma_0, \boldsymbol{\gamma}_1', \boldsymbol{\gamma}_2')$. Standard linear regression methods should be used to select the appropriate $\mathbf{W}$ and $\mathbf{Z}$, and to insure fit of the model to the data.
3. Use the estimated parameters from the first two steps to obtain an estimate of the exposure effect $\hat{\beta}_1 = \boldsymbol{\tau}'\hat{\boldsymbol{\beta}}_{\mathbf{X}}$ where $\boldsymbol{\tau}$ are weights determined below and $\hat{\boldsymbol{\beta}}_{\mathbf{X}}$ are the measurement error-corrected coefficients for each of the surrogates defined as $\hat{\boldsymbol{\beta}}_{\mathbf{X}} = \hat{\Gamma}_1^{-1}\hat{\boldsymbol{\alpha}}_1$ where Γ_1 is a diagonal matrix with the elements of $\boldsymbol{\gamma}_1$ on the diagonal. The weights $\boldsymbol{\tau}$ are determined using generalized least squares theory (Arnold, 1981, p. 202) as $\boldsymbol{\tau}' = (\mathbf{1}'\hat{\Sigma}^{-1}_{\beta_{\mathbf{X}}}\mathbf{1})^{-1}\mathbf{1}'\hat{\Sigma}^{-1}_{\beta_{\mathbf{X}}}$ where $\mathbf{1}$ is an r-vector of 1's and $\hat{\Sigma}_{\beta_{\mathbf{X}}}$ using the multivariate δ-method (Lehmann, 1983, p. 344). With $\mathbf{A}_1$ defined as a diagonal matrix with the elements of $\boldsymbol{\alpha}_1$ on the diagonal, the variance-covariance matrix is given by
$$\hat{\Sigma}_{\boldsymbol{\beta}\mathbf{X}} = \begin{pmatrix} \hat{\Gamma}_1^{-1} \\ -(\hat{\Gamma}_1\hat{\Gamma}_1)^{-1}\hat{\mathbf{A}}_1 \end{pmatrix}' \begin{pmatrix} \hat{\Sigma}_{\boldsymbol{\alpha}_1} & \mathbf{0} \\ \mathbf{0} & \hat{\Sigma}_{\boldsymbol{\gamma}_1} \end{pmatrix} \begin{pmatrix} \hat{\Gamma}_1^{-1} \\ -(\hat{\Gamma}_1\hat{\Gamma}_1)^{-1}\hat{\mathbf{A}}_1 \end{pmatrix}. \tag{4}$$
4. Obtain the estimated corrected coefficients for the intercept, $\hat{\beta}_0 = \hat{\alpha}_0 - \hat{\gamma}_0\hat{\beta}_1 - \hat{\sigma}^2_{X|\mathbf{W},\mathbf{Z}}\hat{\beta}_1^2/2$, and the perfectly measured covariates $\mathbf{Z}$, $\hat{\boldsymbol{\beta}}_2 = \hat{\boldsymbol{\alpha}}_2 - \hat{\beta}_1\hat{\boldsymbol{\gamma}}_2$.
5. Compute the estimated variance–covariance matrix of $\hat{\boldsymbol{\beta}}' = (\hat{\beta}_0, \hat{\beta}_1, \hat{\boldsymbol{\beta}}_2')$ derived by the multivariate δ-method as shown in Appendix B.

If the validation study contains information on outcome (internal validation study), the exposure effect can be estimated within the internal validation study and combined with the proposed estimator, again using an inverse-variance weighted

estimator to minimize the variance of the final summary estimate (Spiegelman et al., 2001). This approach assumes that sampling into the validation study is conditionally independent of $(D, X)|\mathbf{W}, \mathbf{Z}$. User-friendly and fully documented Splus and SAS routines implementing the proposed methods are available (`http://www.hsph.harvard.edu/faculty/spiegelman/multsurr.html`).

4. Application

To illustrate the application of these methods, 1040 (n_1) of the workers in Greaves et al.'s epidemiologic study (1997) and 83 (n_2) of the workers in the validation study (Woskie et al., 1994) who were either working in an area with no direct exposure to MWF or were exposed to synthetic or straight MWF were analyzed. Of the 1811 workers with complete data from the two studies, 236 were in plant 3 and of those not in plant 3, 452 were exposed to soluble MWF, leaving a total of 1123 workers (1040 in the main study and 83 in the validation study). Because the measurement error model estimated in the validation study is assumed to be an estimate of the measurement error model that generated the main study data (the "transportability" assumption described by Carroll et al., 1995, p. 10), we excluded subjects in plant 3 from the main study since exposure was validated only in plants 1 and 2. In addition, we excluded workers in both the main and validation study who were exposed to soluble MWFs. Soluble MWF is, by far, the most common type of MWF in use. It is a hybrid of straight and synthetic MWF. Like straight, soluble includes mineral oils, forming potentially toxic polycyclic aromatic hydrocarbons when heated, and like synthetics, solubles are water-based, thereby permitting growth of mycobacteria and other bioaerosols. Because it combines toxic aspects of each of the other two types, soluble MWFs are difficult to study (Mirer, 2003). For clarity of presentation, we were advised by our epidemiologist (EAE) to exclude workers exposed to this type of MWF in the present analysis.

The effects of metal working fluids (MWF) on the prevalence of wheeze (D), as defined by Greaves et al. as chest sounds wheezy or whistling most days or nights, were estimated. The quantitative exposure measure (X) was the thoracic aerosol fraction in mg/m^3. The surrogates for this exposure ($\mathbf{W}$) were machine type (grinding or not grinding) and MWF type (no direct exposure to fluid, straight MWF or exposure to synthetic MWF). Age, race, plant and smoking status were the perfectly measured covariates ($\mathbf{Z}$) expected to influence the odds of wheeze and, therefore, were considered potential confounders of the (D, X) association. Consistent with the analysis reported in Greaves et al. (1997), the age categories were defined as < 30, 30–39, 40–49, 50+. Workers using straight MWF (29%) and synthetic MWF (37%) had a higher prevalence of wheeze as compared to those using no fluid (20%). Among the validation study workers, the mean thoracic aerosol fraction was higher in those who used straight MWF (0.64 mg/m^3) and synthetic (0.42 mg/m^3) MWF compared to those who had no direct exposure to MWF (0.13 mg/m^3).

The results from fitting the uncorrected logistic regression model are given in Table 1. The uncorrected model results showed a significant increased odds of wheeze for those who were exposed to straight fluid ($p = 0.01$) or synthetic fluid

Table 1
Results from the uncorrected logistic regression for wheeze among the $n_1 = 1040$ main study participants

Variable[a]	$\hat{\boldsymbol{\alpha}}$[b]	$\widehat{\mathrm{SE}}(\hat{\boldsymbol{\alpha}})$	OR (95% CI)	p-value
Intercept	−2.54	0.27		
Surrogates ($\mathbf{W}$)				
Grinding	−0.35	0.32	0.7 (0.4, 1.3)	0.28
Straight MWF	0.50	0.20	1.6 (1.1, 2.4)	0.01
Synthetic MWF	0.62	0.22	1.9 (1.2, 2.9)	0.005
Covariates ($\mathbf{Z}$)				
Plant 2	0.75	0.21	2.1 (1.4, 3.2)	< 0.001
Age (30–39)	−0.11	0.19	0.9 (0.6, 1.3)	0.57
Age (40–49)	−0.18	0.25	0.8 (0.5, 1.4)	0.47
Age (50+)	−0.09	0.26	0.9 (0.5, 1.5)	0.73
Race	0.16	0.20	1.2 (0.8, 1.7)	0.42
Current smoker	1.11	0.16	3.0 (2.2, 4.2)	< 0.001

[a]Reference groups: plant 1, no grinding, assemblers, age < 30, non-white race, not a current smoker.

[b]Estimates for surrogates are based on binary variables. Therefore, the OR compare the odds for wheeze in each surrogate group relative to assembly workers. The mean exposure to thoracic aerosol fraction for straight and synthetic MWF is 0.51 and 0.29 mg/m^3.

Table 2
Results from linear measurement error model of thoracic aerosol fraction in the validation study ($n_2 = 83$)

Variable	$\hat{\gamma}$	$\widehat{\mathrm{SE}}(\hat{\gamma})$	p-value
Intercept	0.15	0.07	0.04
Grinding	0.10	0.07	0.15
Straight	0.50	0.05	< 0.001
Synthetic	0.30	0.06	< 0.001
Plant 2	−0.04	0.08	0.64
Age (30–39)	−0.07	0.06	0.26
Age (40–49)	−0.02	0.07	0.76
Age (50+)	−0.002	0.07	0.99
Race	0.005	0.05	0.91
Current smoker	0.020	0.038	0.60

Table 3
Results from the corrected logistic regression for wheeze

Variable[a]	$\hat{\boldsymbol{\beta}}$	$\widehat{\mathrm{SE}}(\hat{\boldsymbol{\beta}})$	OR (95% CI)	p-value
Intercept	−2.71	0.29		
Exposure (mg/m^3)	1.06	0.38	2.9 (1.4, 6.1)	0.006
Covariates ($\mathbf{Z}$)				
Plant 2	0.78	0.23	2.2 (1.4, 3.4)	< 0.001
Age (30–39)	−0.04	0.20	1.0 (0.7, 1.4)	0.86
Age (40–49)	−0.16	0.26	0.9 (0.5, 1.4)	0.54
Age (50+)	−0.09	0.27	0.9 (0.5, 1.6)	0.74
Race	0.15	0.20	1.2 (0.8, 1.7)	0.45
Current smoker	1.09	0.17	3.0 (2.1, 4.1)	< 0.001

[a]Reference groups: plant 1, no grinding, assemblers, age < 30, non-white race, not a current smoker.

($p = 0.005$), those who were current smokers ($p < 0.001$) and those who worked in plant 2 ($p < 0.001$). The estimated odds ratios (95% CI) were 1.6 (1.1, 2.4) and 1.9 (1.2, 2.9) for straight and synthetic fluid, respectively, relative to those not directly exposed (the assemblers). Since the mean thoracic aerosol fraction (mg/m^3) in those exposed to straight MWF and synthetic MWF was 0.51 and 0.29 units above those with no direct exposure to MWF, respectively, the odds ratio can be interpreted, approximately, as the effect of a 0.51 and 0.29 average increase in exposure to thoracic aerosol fraction in relation to wheeze. Both MWFs were significant predictors of thoracic aerosol fraction (mg/m^3) in the linear regression measurement error model (Table 2) with $\Sigma_{X|\mathbf{W},\mathbf{Z}} = 0.025$. The correlation between the predicted and observed thoracic aerosol fraction of 0.82.

Using the estimates of $\boldsymbol{\alpha}_1$ given in Table 2 and γ_1 given in Table 2, we obtained the vector $\hat{\boldsymbol{\beta}}_{\mathbf{X}} = (-3.54, 0.99, 2.06)$ with standard errors of 4.08 ($p = 0.39$), 0.40 (0.01) and 0.86 ($p = 0.02$), respectively. These estimates were combined using the inverse-variance weights ($\tau = 0.02, 0.86, 0.13$) for grinding, straight MWF and synthetic MWF, respectively. The weight for grinding was small, and therefore, it had little impact on the estimate of β_1. Corrected for measurement error, the estimated odds ratios (95% CI) obtained from the groups of workers exposed to straight and synthetic MWF, respectively, were 2.7 (1.2, 5.9) and 7.9 (1.5, 42.1) per 1 mg/m^3 increase in thoracic aerosol fraction exposure.

After combining the separate effect estimates using the weights ($\boldsymbol{\tau}$), the common aerosol exposure estimate (mg/m^3) was $\hat{\beta}_1 = 1.06$ with an estimated standard error of 0.38 (Table 3). The estimated odds ratio for a 1 mg/m^3 increase in exposure to thoracic particulate matter in relation to wheeze was 2.9 ($p = 0.006$) with a 95% confidence interval of (1.4, 6.1). These results are comparable to Greaves et al. (1997), who concluded that wheeze was associated with current exposure (mg/m^3) to both straight and synthetic MWF. In the present analysis, with correction for measurement error, a single odds ratio summarizes the results in the units of the exposure of interest (mg/m^3) is given as 2.9 per mg/m^3 of MWF exposure. In a simulation to evaluate the performance of the estimator in the setting of this data, the negative percent bias of the proposed estimator was 3.9% and the coverage probability was 95.5%.

Similar estimates of the exposure effect and the corresponding odds ratio were found using the Carroll et al. approach (1995) with $\hat{\beta}_1 = 1.05$ and an estimated asymptotic standard error of 0.37. The estimated odds ratio is 2.9 with a 95% CI (1.4, 5.9). The results from the sensitivity analysis to evaluate the bias and coverage probability were also similar (3.9% negative percent bias and 95.5% coverage probability).

5. A simulation study

Simulation studies were performed to assess the properties of the proposed estimator in small samples and under violations of the rare disease assumption and/or small parameter of measurement error approximation ($\beta_1^2\sigma_{X|W}^2$). The performance was summarized using percent relative bias $\left(100 \times \left(\sum_{b=1}^{2000} \hat{\beta}_{1b}/2000 - \beta_1\right)/\beta_1\right)$, where $\hat{\beta}_{1b}$ is the estimated β_1 from the bth simulated data set and by coverage probability, the percentage of the 2000 generated data sets for which the 95% confidence interval (CI) contained the true parameter value, β_1. In addition, the median correlation between the true exposure (X) and the predicted exposure ($\hat{X}$) in the 2000 generated data sets was computed to indicate the extent of measurement error. For each of the 2000 generated validation and main study samples, the proposed approach was applied to obtain the adjusted exposure effect in each simulation.

For the simulations assessing small sample properties, three sample sizes were considered for the main study: $n_1 = 500, 1000, 2000$. Validation samples sizes considered were $n_2 = 0.05n_1, 0.10n_1, 0.25n_1$. For each of the nine simulations, the exposure effect and measurement error model variance were set equal to that estimated in the motivating example ($\beta_0 = -2$, $\beta_1 = 1.056$ and $\sigma_{X|\mathbf{W}}^2 = 0.025$). The parameters from the measurement error model ($\boldsymbol{\gamma}$) were also fixed at those observed in this example (see Table 2 for these values).

In the simulations assessing the violations of the assumptions, two baseline prevalence of disease ($\beta_0 = (-2, 0)$), two common exposure effects ($\beta_1 = (1.056, 1.758)$) and various values of measurement error were considered ($\sigma_{X|\mathbf{W}}^2 = (0.025, 0.050, 0.090, 0.179, 0.269, 0.359, 0.448)$). These values were considered in order to assess the performance of the estimator under the violation of the rare disease assumption ($\beta_0 = 0$), large exposure effect of X on D ($\beta_1 = 1.758$, corresponds to tripling of the odds ratio observed in the example), and large parameter of small measurement error approximation ($\beta_1^2\sigma_{X|\mathbf{W}}^2$ ranging from 0.027–1.39). The values of $\sigma_{X|\mathbf{W}}^2$ were selected to assess whether the cutoff of $\beta_1^2\sigma_{X|W}^2 < 0.5$ for small measurement error as identified by Kuha (1994) was also appropriate for the proposed estimator. The main and validation study sample sizes were set at 1000 and 100, respectively. As in the small sample simulations, $\boldsymbol{\gamma}$ was fixed at the values observed in the example.

For a particular value of n_1, n_2, β_0, β_1 and $\sigma_{X|\mathbf{W}}^2$, a validation study and main study was generated. To generate the validation study, we sampled $\mathbf{W}$ (machine type and MWF type) jointly with replacement from the validation study stratified by MWF type using the observed allocation scheme. Next n_2 random errors, denoted as $\mathbf{e}$, were generated from a $N(0,\sigma_{X|\mathbf{W}}^2)$. The true exposure values X were computed using the estimated parameters given in Table 2 for γ_0 and $\boldsymbol{\gamma}_1$ from the linear regression model as

$$X = \gamma_0 + \mathbf{W}\boldsymbol{\gamma}_1 + \mathbf{e}. \tag{5}$$

To generate the main study sample, $\mathbf{W}$ was also sampled jointly with replacement from the main study stratified by MWF type. Random errors were generated and true exposure was computed as in (5). Next, D was generated from the Bernoulli distribution with a probability $\Pr(D|X)$ such that $\text{logit}(\Pr(D|X)) = \beta_0 + \beta_1 X$.

5.1. Simulation results

The negative percent bias in the small sample simulations was less than 8.3% in all cases and ranged from −0.3 to 8.3%, even when $n_1 = 500$ and $n_2 = 25$. In addition, the coverage probability was within 2% of the nominal 95% with an expected error range of 1% ($1.96\sqrt{0.95 \times 0.05/2000}$) and ranged from 95.0 to 95.9%.

The results from the violation of assumption simulations were more varied and are presented in Table 4 . The negative percent bias increases and coverage probability decreased as the parameter of small measurement error approximation ($\beta_1^2\sigma_{X|\mathbf{W}}^2$) increased. The negative percent bias was slightly higher and the coverage probability was slightly lower for the 'common disease' scenario ($\beta_0 = 0$) compared to the 'rare disease' scenario ($\beta_0 = -2$).

The cutoff proposed by Kuha (1994) cannot be used globally for the proposed estimator. We observed that is some scenarios studied by simulation where $\beta_1^2\sigma_{X|\mathbf{W}}^2 < 0.5$, the negative percent bias is as high as 42.8% and the coverage

Table 4
Simulation results exploring the violation of assumptions with allocation ratio (0.6, 0.2, 0.2) for assemblers, straight and synthetic workers $n_1 = 1000$, $n_2 = 100$, 2000 simulations

β_0	β_1	$\sigma^2_{X\|W}$	$\beta_1^2\sigma^2_{X\|W,Z}$	% Bias	Cov. Prob. (%)	Median Corr $(X, \hat{X})$
−2	1.056	0.025	0.027	−3.9	95.5	0.81
		0.050	0.056	−6.5	95.2	0.70
		0.090	0.100	−10.5	94.9	0.60
		0.179	0.200	−18.7	91.6	0.48
		0.269	0.300	−26.5	86.8	0.41
		0.359	0.400	−33.5	81.3	0.38
		0.448	0.500	−39.7	76.2	0.35
	1.758	0.025	0.077	−4.3	94.5	0.81
		0.050	0.155	−8.0	92.4	0.70
		0.090	0.275	−14.2	88.6	0.60
		0.179	0.553	−26.3	74.1	0.48
		0.269	0.831	−36.8	60.9	0.41
		0.359	1.110	−44.8	49.3	0.38
		0.448	1.385	−51.5	41.4	0.35
0	1.056	0.025	0.027	−4.1	94.9	0.81
		0.050	0.056	−8.9	94.7	0.70
		0.090	0.100	−16.1	85.6	0.60
		0.179	0.200	−21.3	87.3	0.48
		0.269	0.300	−30.3	79.4	0.41
		0.359	0.400	−37.0	72.7	0.38
		0.448	0.500	−42.8	65.3	0.35
	1.758	0.025	0.077	−4.1	94.9	0.81
		0.050	0.155	−8.9	91.7	0.70
		0.090	0.275	−16.1	85.6	0.60
		0.179	0.553	−28.8	68.0	0.48
		0.269	0.831	−39.3	52.6	0.41
		0.359	1.110	−47.3	43.6	0.38
		0.448	1.385	−53.2	36.8	0.35

probability is as low as 65.3. However, the proposed estimator performed well with a negative percent bias less than 10% and coverage probability within 1% of the nominal 95% when $\beta_1^2\sigma^2_{X|\mathbf{W}} \leqslant 0.077$ for both moderate and strong exposure effects ($\beta_1 = 1.056, 1.758$) and for both rare and common disease scenarios ($\beta_0 = -2, 0$). When $\sigma^2_{X|\mathbf{W}} \leqslant 0.05$, the median correlation between X and $\hat{X}$ was greater than or equal to 0.70. It also performed well with median correlations as low as 0.60 with $\beta_0 = -2$ and $\beta_1 = 1.056$. In the motivating example the correlation was 0.82 and $\hat{\beta}_1^2\hat{\sigma}^2_{X|\mathbf{W},\mathbf{Z}} = 0.028$.

Since a global cut point could not be established, an option is included in our publicly available Splus function and SAS macro to perform a sensitivity analysis over a range of values of β_1 to evaluate the percent bias and the coverage probability of the estimator for a particular data set. The data is generated as described above with the perfectly measured covariates sampled jointly with the surrogates from the main study/validation study. Because the true value of β_1 is unknown in an applications setting, the properties of the estimator are evaluated for bias and coverage probability at the estimated value of β_1 and $\pm 0.25\beta_1$.

6. Asymptotic relative efficiency (ARE) evaluation

6.1. Carroll et al. (1995) approach

The approach proposed by Carroll et al. (1995) to adjust for estimated exposure values in the health effects model involves solving a non-standard set of estimating equations derived from a Taylor series approximation of the mean and variance functions. The method simplifies considerably with homoscedastic variance and an assumption that the

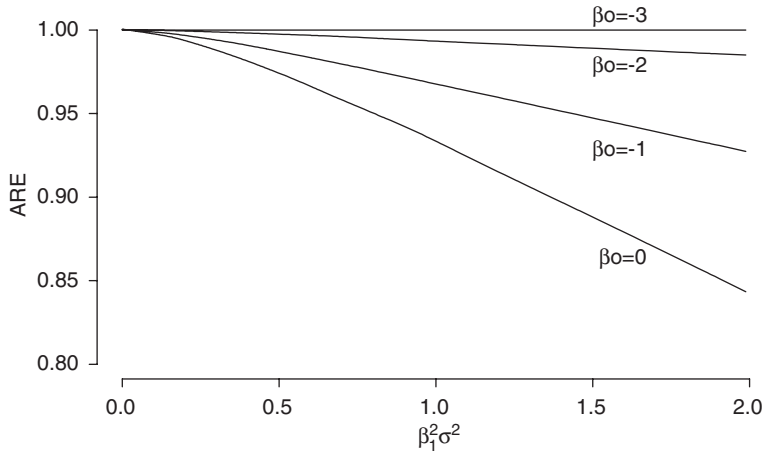

Fig. 1. Asymptotic relative efficiency of the Carroll et al. estimator relative to the proposed estimator across the parameter of small measurement error approximation. Results are from a grid search with weights outside the interval (0.4, 0.6) and are combined across all values of γ.

measurement error variance is small. Then, the validation data (X, $\mathbf{W}$, $\mathbf{Z}$) can be used to estimate the measurement error coefficients (γ) and estimate the predicted true exposure variables for all main study subjects. The unobserved true variables X are replaced with their predicted values in the logistic regression model for the subjects in the main study and a standard analysis is run to obtain the parameter estimates. To obtain the adjusted standard errors, a bootstrap method or the estimating equations approach can be used. The asymptotic variance of $\hat{\boldsymbol{\beta}}$ is derived in Appendix C using the model-based expectation method (Carroll et al., 1995, p. 263) approach which uses an underlying model to determine the covariance exactly and then substitutes the estimated parameters.

6.2. ARE results

The ARE of the estimator given by Carroll et al. (1995) relative to the proposed estimator for β_1 was evaluated numerically over a grid centered around the parameters estimated in the data from the motivating example, for the case with two surrogates ($\mathbf{W}$ = straight MWF and synthetic MWF) for one exposure (X = thoracic aerosol fraction in mg/m^3) and with no perfectly measured covariates. In this evaluation, the data from the example are used (that is, the mean and variance of X and $\mathbf{W}$) are held constant), the intercept for the measurement error model is fixed at that observed in the example ($\gamma_0 = 0.127$) and the parameters for the surrogates in the measurement error model (γ_{11},γ_{12}), the disease prevalence (β_0) and the exposure effects (β_1) are varied.

The boundaries of the grid search were selected based on the motivating example. The parameters for the two surrogates for exposure in the measurement error model ranged from 0.1 to 1.5, specifically, γ_{11}=γ_{12}=(0.1, 0.25, 0.5, 1, 1.5). Four disease prevalence rates were considered: $\beta_0 = (-3, -2, -1, 0)$ and 22 values of β_1 ranging from 0.05 to 3.0. ARE of the proposed estimator relative to the Carroll et al. (1995) estimator was computed for the 1123 points of this grid. The proposed estimator was shown in all cases considered to have a variance less than or equal to the Carroll et al. estimator. The results show that in the cases with τ_1 and τ_2 between 0.4 and 0.6, the ARE is approximately equal to one for all values of γ_{11}, γ_{12}, β_0 and $\beta_1^2\sigma^2_{X|\mathbf{W},\mathbf{Z}} < 2$. This suggests that with approximately equal weights the ARE is equal to 1. When the weights are not both in (0.4, 0.6), the ARE (Fig. 1) is dependent upon the prevalence of the disease (β_0), the effect of the surrogates on exposure in the measurement error model (γ_1), and the parameter of small measurement error approximation ($\beta_1^2\sigma^2_{X|\mathbf{W},\mathbf{Z}}$). In these cases, the ARE gets smaller (< 1) as we move away from the rare disease assumption and away from the small measurement error approximation.

7. Discussion

In this paper, a regression calibration approach is proposed that provides a single estimate of the exposure effect on a health outcome by combining the parameter estimates for each surrogate (such as job area, type of work) using

inverse variance weights which minimize the variance of the summary estimator. Each of these single estimates of the exposure effect is adjusted for exposure measurement error. This approach assumes a linear measurement error model and a logistic regression model for the health effects model. If the variance of $\hat{\boldsymbol{\beta}}_X$ is estimated adequately, this estimator has the smallest variance among all linear combinations of the elements of $\hat{\boldsymbol{\beta}}_X$. The estimator is approximately consistent when the disease is rare and the measurement error is normally distributed, or if $\beta_1^2\sigma^2_{X|\mathbf{W},\mathbf{Z}}$ is small, and if, as usual, the models are correctly specified. The simulation studies showed that the proposed estimator performed well under the circumstances of the motivating example, with small sample sizes, in regions of the parameter space where β_1 and $\sigma^2_{X|\mathbf{W}}$ are likely to be in most practical situations, under violations of the rare disease assumption and for correlations between X and $\hat{X}$ greater than or equal to 0.70. Over a wide grid search centered around the parameters estimated in the motivating example, the variance of the proposed estimator was less than that of the Carroll et al. (1995) estimator with the largest difference occurring when the weights ($\boldsymbol{\tau}$) were not equal. A uniform cutoff similar to Kuha (1994) could not be determined. Therefore, the authors include an option within the S-plus code and SAS macro to run a sensitivity analysis with parameters defined by the data set at hand.

Alternative approaches include likelihood methods, latent variable models and latent trait models. Likelihood-based methods for estimation and inference about β are more computationally complex than the regression calibration approaches proposed. Regression calibration for logistic regression is limited by assumptions, some of which are obviated by turning to maximum likelihood methods. In particular, regression calibration assumptions that require the outcome of interest to be rare, the measurement error and true exposure effect to be small and the relationship between the surrogates and the true exposure to be linear and homoscedastic are no longer necessary. However, in situations where these assumptions are reasonable, the regression calibration approach is preferred because it is in line with the investigators' framework of obtaining initial estimates and adjusting for measurement error. In addition, the implementation of customized non-standard likelihood for routine data analysis uses remains an insurmountable barrier in most applied settings in epidemiology, and occupational and environmental health research.

It is suggested by the work of Spiegelman et al. (2005), Kipnis et al. (1999), and Kaaks et al. (1994), all of whom consider forms of latent variable or structural equation models, that unless one of the surrogates is unbiased, the model which relates the mean of the true exposure to the surrogates in not identified, either when full multivariate normality is assumed or when assumptions are made only about the models for the first two moments, at least when errors in the surrogates are correlated. Since the surrogates are not self-reported, as in the nutritional epidemiology settings, it might be reasonable in this setting to assume that the errors are uncorrelated, in which case a latent variable model such as that given by Spiegelman et al. (2005) could be fit to the validation study data to allow a full data maximum likelihood or pseudo-likelihood fit to the data for estimation and inference about β_1. The latent trait literature (e.g. Lord and Novik, 1968; Cohen and Jiang, 1999) may also have a contribution to make to this problem, having previously been applied mostly to applications in psychology. Further research is needed to elaborate these methods and compare them with respect to statistical and computational efficiency with existing alternatives.

Estimating the health effects from occupational exposures from epidemiological data and exposure assessment data remains a challenging problem. In many cases, the estimated average individual values are used directly in the model relating exposure to the health outcome without any adjustment for the fact that the exposure values are estimated. We provide a regression calibration method to combine multiple surrogates for one exposure. The proposed approach is valid if the disease is rare or when measurement error is small and/or when the dependence of D on X is not too strong. It performs well in the small sample setting and is not sensitive to departures from the rare disease assumption as long as the parameter of small measurement error approximation is small. Programs in Splus and SAS implementing these methods are available (`http://www.hsph.harvard.edu/faculty/spiegelman/multsurr.html`). Use of this approach should improve the estimated exposure effects on health outcomes in epidemiological studies where multiple surrogates describe one exposure.

Acknowledgments

The authors thank the members of the Measurement Error Working Group at the Harvard School of Public Health and Ruifeng Li for their valuable comments and suggestions. This research was supported by the following grants: NIEHS ES09411-03, NIOSH RO1 OH03489, NIEHS RO1 ES007036 and NIEHS 2P30ES00002.

Appendix A. Derivation of estimator using a Taylor series expansion

In cases where the rare disease assumption is not appropriate, an approximation of $\Pr(D=1|\mathbf{W},\mathbf{Z})$ can be found by substituting the second-order Taylor series expansion of $\Pr(D=1|X,\mathbf{Z})$ about $\mu_{X|\mathbf{W},\mathbf{Z}}$ in $\Pr(D=1|\mathbf{W},\mathbf{Z}) \approx \int \Pr(D=1|X,\mathbf{Z}) f_{X|\mathbf{W},\mathbf{Z}}(X|\mathbf{W},\mathbf{Z})\,\mathrm{d}X$. The following approximation is found, which is similar to that in Rosner et al. (1989) and Kuha (1994),

$$\begin{aligned}&\Pr(D=1|\mathbf{W},\mathbf{Z})\\ &\quad\approx \frac{\exp[\beta_0+\beta_1\mu_{X|\mathbf{W},\mathbf{Z}}+\mathbf{Z}\boldsymbol{\beta}_2]}{1+\exp[\beta_0+\beta_1\mu_{X|\mathbf{W},\mathbf{Z}}+\mathbf{Z}\boldsymbol{\beta}_2]}\\ &\qquad+\frac{\exp[\beta_0+\beta_1\mu_{X|\mathbf{W},\mathbf{Z}}+\mathbf{Z}\boldsymbol{\beta}_2](1-\exp[\beta_0+\beta_1\mu_{X|\mathbf{W},\mathbf{Z}}+\mathbf{Z}]\boldsymbol{\beta}_2)\beta_1^2\sigma^2_{X|\mathbf{W},\mathbf{Z}}}{2(1+\exp[\beta_0+\beta_1\mu_{X|\mathbf{W},\mathbf{Z}}+\mathbf{Z}\boldsymbol{\beta}_2])^3}.\end{aligned}$$

After taking the expectation of the expansion, the linear term is equal to zero. If $\beta_1^2\sigma^2_{X|\mathbf{W},\mathbf{Z}}$ is close to zero, the second term disappears and $\Pr(D=1|\mathbf{W},\mathbf{Z})$ is approximated by

$$\Pr(D=1|\mathbf{W},\mathbf{Z}) \approx \frac{\exp[\beta_0+\beta_1\mu_{X|\mathbf{W},\mathbf{Z}}+\mathbf{Z}\boldsymbol{\beta}_2]}{1+\exp[\beta_0+\beta_1\mu_{X|\mathbf{W},\mathbf{Z}}+\mathbf{Z}\boldsymbol{\beta}_2]}.$$

Substituting the expression for the mean of X given in (3), the estimates of β_1 and β_2 are the same as those derived in Section 2. Assuming that $\beta_1^2\sigma^2_{X|\mathbf{W},\mathbf{Z}}$ is close to zero, the estimate of β_0 is also the same as given in Section 2.

Appendix B. Variance of proposed estimator

The estimated $(s+2)\times(s+2)$ variance-covariance matrix of $\hat{\boldsymbol{\beta}}'=(\hat{\beta}_0,\hat{\beta}_1,\hat{\boldsymbol{\beta}}_2')$ is determined using the multivariate δ-method as follows:

$$\begin{aligned}\hat{\Sigma}_{\boldsymbol{\beta}} &= \left(\frac{\partial\boldsymbol{\beta}}{\partial(\boldsymbol{\alpha},\gamma,\sigma^2_{X|\mathbf{W},\mathbf{Z}})}\right)'_{(\hat{\boldsymbol{\alpha}},\hat{\gamma},\hat{\sigma}^2_{X|\mathbf{W},\mathbf{Z}})}\begin{pmatrix}\hat{\Sigma}_{\boldsymbol{\alpha}} & \mathbf{0} & 0\\ \mathbf{0} & \hat{\Sigma}_{\gamma} & 0\\ \mathbf{0} & \mathbf{0} & \hat{\Sigma}_{\sigma^2_{X|\mathbf{W},\mathbf{Z}}}\end{pmatrix}\\ &\quad\times\left(\frac{\partial\boldsymbol{\beta}}{\partial(\boldsymbol{\alpha},\gamma,\sigma^2_{X|\mathbf{W},\mathbf{Z}})}\right)_{(\hat{\boldsymbol{\alpha}},\hat{\gamma},\hat{\sigma}^2_{X|\mathbf{W},\mathbf{Z}})},\end{aligned}$$

where $\hat{\Sigma}_{\gamma}=(U'U)^{-1}\hat{\sigma}^2_{X|\mathbf{W},\mathbf{Z}}$, $U=(1,W,Z)$, and $\hat{\Sigma}_{\boldsymbol{\alpha}}=I(\boldsymbol{\alpha})^{-1}$ where $I(\boldsymbol{\alpha})$ is the information matrix from the logistic regression model. The variance of $\hat{\sigma}^2_{X|\mathbf{W},\mathbf{Z}}$ is estimated under normality to be $\hat{\Sigma}_{\sigma^2_{X|\mathbf{W},\mathbf{Z}}}=2\hat{\sigma}^4_{X|\mathbf{W},\mathbf{Z}}/(n_2-(r+s+1))$ (Seber, 1977, p. 52). The $(2r+2s+3)\times(s+2)$ derivative matrix is given by

$$\begin{aligned}&\frac{\partial\boldsymbol{\beta}'}{\partial(\boldsymbol{\alpha},\gamma,\sigma^2_{X|\mathbf{W},\mathbf{Z}})}\\ &\quad=\begin{pmatrix}1 & 0 & \mathbf{0}_{1\times s}\\ -(\gamma_0+\sigma^2_{X|\mathbf{W},\mathbf{Z}}\beta_1)\Gamma_1^{-1}\boldsymbol{\tau} & \Gamma_1^{-1}\boldsymbol{\tau} & -\Gamma_1^{-1}\boldsymbol{\tau}\boldsymbol{\gamma}_2'\\ \mathbf{0}_{s\times 1} & \mathbf{0}_{s\times 1} & \mathbf{I}_{s\times s}\\ -\beta_1 & 0 & \mathbf{0}_{1\times s}\\ (\gamma_0+\sigma^2_{X|\mathbf{W},\mathbf{Z}}\beta_1)\mathrm{A}_1(\Gamma_1\Gamma_1)^{-1}\boldsymbol{\tau} & -\mathrm{A}_1(\Gamma_1\Gamma_1)^{-1}\boldsymbol{\tau} & \mathrm{A}_1(\Gamma_1\Gamma_1)^{-1}\boldsymbol{\tau}\boldsymbol{\gamma}_2'\\ \mathbf{0}_{s\times 1} & \mathbf{0}_{s\times 1} & -\beta_1\mathbf{I}_{s\times s}\\ -\beta_1^2/2 & 0 & 0_{1\times s}\end{pmatrix}.\end{aligned}$$

Appendix C. Derivation of Carroll et al. (1995) variance

The asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ using the Carroll et al. (1995) approach is determined by stacking the estimating equations for $\boldsymbol{\theta} = (\boldsymbol{\gamma}, \boldsymbol{\beta})$ as described in Carrol et al. (1995, pp. 263–268). The model-based method is used to estimate the components of the covariance matrix. Here, we consider the case with no perfectly measured covariates. Denoting $n = n_1 + n_2$, the estimating equations are:

$$\Psi(\boldsymbol{\gamma}) = \frac{1}{n}\mathbf{U}' I_v (X - \mathbf{U}\boldsymbol{\gamma}) = 0$$

and

$$\Psi(\boldsymbol{\beta}) = \frac{1}{n}\boldsymbol{\gamma}'_{\text{mat}}\mathbf{U}' I_m (D - \mu) = 0$$

where $\mathbf{U} = (\mathbf{1}, \mathbf{W}, \mathbf{Z})$, $\mathbf{W}$ is $n \times r$, $\mathbf{Z}$ is $n \times s$ and I_m, I_v are $n \times n$ diagonal matrices with elements equal to 1 if subjects are in the main study for I_m and elements equal to 0 if subjects are in the validation study for I_v. In the estimating equation for $\boldsymbol{\beta}$,

$$\mu = g^{-1}(U\boldsymbol{\gamma}_{\text{mat}}\boldsymbol{\beta}), \quad g^{-1}(\cdot) = \frac{\exp(\cdot)}{1 + \exp(\cdot)}$$

and

$$\boldsymbol{\gamma}_{\text{mat}} = \begin{pmatrix} 1 & \gamma_0 & \mathbf{0}_s \\ \mathbf{0}'_r & \boldsymbol{\gamma}_1 & \mathbf{0}_{r\times s} \\ \mathbf{0}'_s & \boldsymbol{\gamma}_2 & \mathbf{I}_{s\times s} \end{pmatrix}$$

where $\mathbf{0}_r$ is an r-vector of zeros, $\mathbf{0}_s$ is an s-vector of zeros, $\mathbf{0}_{r\times s}$ is a $s \times r$-matrix of zeros, and $\mathbf{I}_{s\times s}$ is an identity matrix. Following A.8–A.9 of Carroll et al. (1995, p. 263), the covariance matrix of $\boldsymbol{\theta}$ is $\frac{1}{n}\mathrm{A}_n^{-1}\mathrm{B}_n(\mathrm{A}_n^{-1})'$. The model-based estimator uses

$$\hat{\mathrm{A}}_n(\hat{\boldsymbol{\theta}}) = \frac{1}{n}\begin{pmatrix} -U'\mathbf{I}_v U & \mathbf{0}_{r+s+1,s+2} \\ -\hat{\beta}_1(\hat{\boldsymbol{\gamma}}'_{\text{mat}}U'\mathbf{I}_m\hat{\mathbf{V}}(D|X)U) & -(\hat{\boldsymbol{\gamma}}'_{\text{mat}}U'\mathbf{I}_m\hat{\mathbf{V}}(D|X)U\hat{\boldsymbol{\gamma}}_{\text{mat}}) \end{pmatrix}$$

and

$$\hat{\mathrm{B}}_n(\hat{\boldsymbol{\theta}}) = \frac{1}{n}\begin{pmatrix} U'\mathbf{I}_v U\hat{\sigma}^2_{X|\mathbf{W},\mathbf{Z}} & \mathbf{0}_{r+s+1,s+2} \\ \mathbf{0}'_{r+s+1,s+2} & (\hat{\boldsymbol{\gamma}}'_{\text{mat}}U'\mathbf{I}_m\hat{\mathbf{V}}(D|X)U\hat{\boldsymbol{\gamma}}_{\text{mat}}) \end{pmatrix}$$

where $\mathbf{V}(D|X)$ is a $n \times n$ diagonal matrix with elements $\mu_i * (1 - \mu_i)$ for $i = 1, 2, \ldots, n$ and $\mathbf{0}_{r+s+1,s+2}$ is a $(r+s+1) \times (s+2)$ matrix of zeros. The asymptotic variance of $\hat{\boldsymbol{\beta}}$ is the lower right sub-matrix of $(1/n)\hat{\mathrm{A}}_n^{-1}\hat{\mathrm{B}}_n(\hat{\mathrm{A}}_n^{-1})'$.

References

Armstrong, B.G., 1990. The effects of measurement errors in relative risk regressions. Amer. J. Epidemiol. 132, 1176–1184.

Arnold, S.F., 1981. The Theory of Linear Models and Multivariate Analysis. Wiley, New York.

Belsey, D.A., Kuh, E., Welsch, R.E., 1980. Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. Wiley, New York.

Carroll, R.J., Stefanski, L.A., 1990. Approximate quasi-likelihood estimation in models with surrogate predictors. J. Amer. Statist. Assoc. 85, 652–663.

Carroll, R.J., Ruppert, D., Stefanski, L.A., 1995. Measurement Error in Nonlinear Models. Chapman & Hall, New York.

Cohen, R.J, Jiang, T., 1999. Comparison of partially measured latent traits across nominal subgroups. J. Amer. Statist. Assoc. 94, 1035–1044.

Eisen, E.A., Tolbert, P.E., Monson, R.R., Smith, T.J., 1992. Mortality studies of machining fluid exposure in the automobile industry. I. A standardized mortality ratio analysis. Amer. J. Ind. Med. 22, 809–824.

Greaves, I.A., Eisen, E.A., Smith, T.J., Pothier, L.J., Kreibel, D., Woskie, S.R., Kennedy, S.M., Shalat, S., Monson, R.R., 1997. Respiratory health of automobile workers exposed to metal-working fluid aerosols: respiratory symptoms. Amer. J. Ind. Med. 32, 450–459.

Hallock, M.F., Smith, T.J., Woskie, S.R., Hammond, S.K., 1994. Estimation of historical exposures to machining fluids in the automotive industry. Amer. J. Ind. Med. 26, 621–634.

Kaaks, R., Riboli, E., Esteve, J., Van Kappel, A., Vab Staveren, W., 1994. Estimating the accuracy of dietary questionnaire assessments: validation in terms of structural equation models. Statist. Med. 13, 127–142.

Kipnis, V., Carroll, R.J., Freedman, L.S., Li, L., 1999. Implications of a new dietary measurement error model for estimation of relative risk: application to four calibration studies. Amer. J. Epidemiol. 150, 642–651.

Kuha, J., 1994. Corrections for exposure measurement error in logistic regression models with an application to nutritional data. Statist. Med. 13, 1135–1148.

Lehmann, E.L., 1983. Theory of Point Estimation. Wiley, New York.

Lee, L.F., Sepanski, J.H., 1995. Estimation of linear and nonlinear errors-in-variables models using validation data. J. Amer. Statist. Assoc. 90, 130–140.

Lord, F.M., Novik, M.R., 1968. Statistical Theories of Mental Test Scores. Addison-Wesley, Reading, MA.

Lyles, R.H., Kupper, L.L., 1997. A detailed evaluation of adjustment methods for multiplicative measurement error in linear regression with applications in occupational epidemiology. Biometrics 53, 1008–1025.

Mirer, F., 2003. Updated epidemiology of workers exposed to metalworking fluids provides sufficient evidence for carcinogenicity. Appl. Occup. Environ. Hyg. 18, 902–912.

Rosner, B., Willett, W.C., Spiegelman, D., 1989. Correction of logistic relative risk estimates for non-random measurement error. Statist. Med. 8, 1051–1069.

Rosner, B., Spiegelman, D., 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. Amer. J. Epidemiol. 132, 734–745.

Rosner, B., Spiegelman, D., Willett, W.C., 1992. Correction of logistic regression relative risk estimates and confidence intervals for random within-person measurement error. Amer. J. Epidemiol. 136, 1400–1413.

Seber, G.A.F, 1977. Linear Regression Analysis. Wiley, New York.

Smith, T.J., 1987. Exposure assessment for occupational epidemiology. Amer. J. Ind. Med. 12, 249–268.

Spiegelman, D., Carroll, R., Kipnis, V., 2001. Efficient regression calibration for logistic regression in main study/internal validation study designs with an imperfect reference instrument. Statist. Med. 10, 139–160.

Spiegelman, D., Zhao, B., Kim, J., 2005. Correlated errors in biased surrogates: study designs and methods for measurement error correction. Statist. Med. 24, 1657–1682.

Thomas, D., Stram, D., Dwyer, J., 1993. Exposure measurement error: influence on exposure-disease relationships and methods of correction. Annu. Rev. Public Health 14, 69–83.

Thurston, S.W., Spiegelman, D., Ruppert, D., 2003. Equivalence of regression calibration methods in main study/external validation study designs. J. Statist. Planning Inference 113, 527–539.

Tolbert, P.E., Eisen, E.A., Pothier, L.J., Monson, R.R., Hallock, M.F., Smith, T.J., 1992. Mortality studies of machining fluid exposure in the automobile industry. II. Risks associated with specific fluid types. Scand. J. Work Environ. Health 18, 351–360.

Woskie, S.R., Smith, T.J., Hallock, M.F., Hammond, S.K., Rosenthal, F., Eisen, E.A., Kreibel, D., Greaves, I.A., 1994. Size-selective pulmonary dose indices for metal-working fluid aerosols in machining and grinding operations in the automobile manufacturing industry. Amer. Ind. Hyg. Assoc. 55, 20–29.