Yi and He *Journal of Statistical Distributions and Applications* (2017) 4:16
DOI 10.1186/s40488-017-0069-0

Journal of Statistical Distributions and Applications

METHODOLOGY Open Access

Analysis of case-control data with interacting misclassified covariates

Grace Y. Yi[1*] and Wenqing He[2]

*Correspondence:
yyi@uwaterloo.ca
[1]Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Canada
Full list of author information is available at the end of the article

Abstract

Case-control studies are important and useful methods for studying health outcomes and many methods have been developed for analyzing case-control data. Those methods, however, are vulnerable to mismeasurement of variables; biased results are often produced if such a feature is ignored. In this paper, we develop an inference method for handling case-control data with interacting misclassified covariates. We use the prospective logistic regression model to feature the development of the disease. To characterize the misclassification process, we consider a practical situation where replicated measurements of error-prone covariates are available. Our work is motivated in part by a breast cancer case-control study where two binary covariates are subject to misclassification. Extensions to other settings are outlined.

Keywords: Case-control study, Interaction term, Misclassification, Prospective logistic regression, Replicated measurements

1 Introduction

Case-control studies are important and useful methods for studying rare health outcomes, such as rare diseases. They do not require us to follow up a large number of subjects over a long period of time. The primary purpose of a case-control study is to investigate how risk factors are associated with the disease incidence, and the study typically involves the comparison of cases (i.e., diseased individuals) with controls (i.e., disease-free individuals).

Various statistical analysis methods for case-control data have been developed in the literature (e.g., Prentice and Pyke 1979; Breslow and Cain 1988; Breslow and Day 1980; Schlesselman 1982). Those methods are, however, vulnerable to mismeasurement of variables that commonly accompanies case-control studies. It has been well documented that ignoring mismeasurement effects in the analysis often yields seriously biased results (e.g., Gustafson et al. 2001; Yi 2017). For instance, Bross (1954) examined misclassification effects on hypothesis testing with 2×2 tables. He commented that misclassification may present a more serious problem in the case of estimates than in the case of significance tests.

In the literature, many methods have been proposed to address mismeasurement effects (e.g., Armstrong et al. 1989; Carroll et al. 1995; Forbes and Santner 1995; Roeder et al. 2005). In particular, with misclassifiction present in discrete covariates or exposure variables, various authors explored strategies for accommodating misclassification effects. To

name a few, Marinos et al. (1995) studied case-control data with non-differential misclassification. Morrissey and Spiegelaman (1999) and Lyles (2002) discussed adjustment methods for exposure misclassification in case-control studies where a validation sample is available. Chu et al. (2009) presented a likelihood-based approach for case-control studies with multiple non-gold standard exposure assessments. Under the Bayesian framework, Prescott and Garthwaite (2005) proposed methods for analyzing matched case-control studies in which a binary exposure variable is subject to misclassification. Tang et al. (2013) considered the case where both disease and exposure are subject to misclassification and developed misclassification adjustment methods by utilizing validation data. Mak et al. (2015) studied sensitivity analysis in case-control studies subject to exposure misclassification.

A common feature of those methods is that error-prone variables enter models separately and no interactions among error-contaminated covariates are considered. In case-control studies, however, error-involved covariates may interactively influence the development of diseases. Incorporating such a feature in the analysis is imperative. Zhang et al. (2008) developed an inference method to account for interacting covariates with misclassification. Their method is applicable only for the case where a validation subsample is available for determining the misclassification probabilities.

In many circumstances, a validation sample is impossible to be collected for various reasons. Some variables may be too expensive or time consuming to be measured precisely (e.g., Carroll et al. 1993). Some variables can never be measured precisely due to its nature. For instance, blood pressure entering a model usually refers to its long term average, and there is no way to obtain this value precisely; exposure amount to a hazard condition, such as radiation, is difficult to be measured accurately. In such situations, a validation sample with precise measurements of the variables is not available. However, surrogate measurements of those variables can sometimes be repeatedly collected in application. In this paper, we consider such a setting where replicated measurements of error-prone covariates are available and develop inference methods with interacting error-prone covariates taken into account. Our development is particularly cast into the framework of the prospective logistic regression model with misclassified binary covariates.

Our research is partially motivated by the case-control study on breast cancer discussed by Duffy et al. (1989). In this study, 451 breast cancer cases were compared with the same number of controls with respect to the error-prone risk factors: alcohol consumption and smoking, where alcohol consumption is defined as a binary variable by a threshold of 9.3 g ethanol/day, and the smoking variable is dichotomized by comparing the product of cigarettes smoked per day and years of smoking to 300. In addition, an independent study on 100 women was available, where repeated measurements of alcohol consumption and smoking were collected for those women on two occasions. It is interesting to study how smoking and alcohol use may be associated with the risk of developing breast cancer and whether or not those two factors may be interacting in explaining the development of breast cancer. A detailed analysis of such data is presented in Section 5.

The remainder of this paper is organized as follows. Section 2 outlines the notation and model setup. In Section 3 we explore the misclassification effects. In Section 4, we develop inferential procedures to accommodate misclassification effects with the availability of replicated measurements of error-contaminated covariates. Analysis of the motivating data with the proposed method is reported in Section 5, together with simulation studies

which demonstrate the performance of our method. The manuscript is concluded with discussion and extensions.

2 Notation and framework

Let Y be the binary outcome variable, taking value 1 if a subject is a case and value 0 otherwise. Let X_a and X_s be two binary covariates taking value 0 or 1, such as alcohol and smoking statuses in the motivating example. For $i, j, k = 0, 1$, let

$$p_{ijk} = P\left(X_a = j, X_s = k | Y = i\right)$$

be the conditional probability for the case or control. Let ψ_{jk} be the odds ratio for cases versus controls with $\left(X_a = j, X_s = k\right)$ compared with the baseline category $(X_a = 0, X_s = 0)$:

$$\psi_{jk} = \frac{p_{000}p_{1jk}}{p_{100}p_{0jk}}$$

for $(j, k) \neq (0, 0)$. Define

$$\psi = \frac{\psi_{11}}{\psi_{01}\psi_{10}}.$$

This measure can be used to indicate the association between the two binary covariates, which is classified by the subpopulations of cases and controls. The measure ψ is defined from the *retrospective* sampling viewpoint which directly reflects the feature of case-control designs. Equivalently, this measure has an equally interpretive feature in a prospective regression model.

Consider the prospective logistic regression model with an interaction term between X_a and X_s:

$$\log\left\{\frac{P(Y = 1|X_a, X_s)}{P(Y = 0|X_a, X_s)}\right\} = \beta_0 + \beta_a X_a + \beta_s X_s + \beta_{as} X_a X_s, \tag{1}$$

where the β_0, β_a, β_s and β_{as} are the regression parameters. These parameters can be expressed in terms of the odds ratios defined for the retrospective sampling framework:

$$\beta_a = \log\psi_{10},\ \beta_s = \log\psi_{01},\ \text{and}\ \beta_{as} = \log\psi. \tag{2}$$

As pointed out by Prentice and Pyke (1979), the baseline parameter β_0 is not estimable from *retrospectively* collected data unless the prevalence $P(Y = 1)$ is known; the coefficients $(\beta_a, \beta_s, \beta_{as})$, or the odds ratios ψ_{jk}, however, is estimable from case-control data that are collected retrospectively.

We now elaborate on estimation procedures. For $i, j, k = 0, 1$, let N_{ijk} represent the number of subjects with $(Y = i, X_a = j, X_s = k)$, and let $(n_{i00}, n_{i10}, n_{i01}, n_{i11})^T$ be a realization of the random vector $N_i = (N_{i00}, N_{i10}, N_{i01}, N_{i11})^T$. Let n_0 and n_1 be the total number of controls and cases in the study, respectively. With the retrospective sampling scheme for case-control studies, these totals are treated as fixed, and it is often plausible to use multinomial distributions to independently characterize the cell counts for the control and case populations. Namely, N_0 and N_1 are assumed to be independent and marginally follow a multinomial distribution with $N_i \sim \text{Multinomial}(n_i, p_i)$, where $p_i = (p_{i00}, p_{i10}, p_{i01}, p_{i11})$ and $\sum_{j,k=0}^{1} p_{ijk} = 1$ and $n_i = \sum_{j,k=0}^{1} n_{ijk}$ for $i = 0, 1$.

These distributional assumptions immediately allow us to write out the likelihood function for the cell probabilities p_{ijk}, ignoring the normalizing constant,

$$L = \prod_{i=0}^{1}\prod_{j=0}^{1}\prod_{k=0}^{1} p_{ijk}^{n_{ijk}}. \tag{3}$$

In combination with the constraint $\sum_{j,k} p_{ijk} = 1$ for a given i, maximizing (3) with respect to the cell probabilities leads to the maximum likelihood estimator for the cell probabilities:

$$\widehat{p}_{ijk} = \frac{n_{ijk}}{n_i} \quad \text{for } i, j, k = 0, 1.$$

Then the invariance of maximum likelihood estimators gives us an estimate of ψ_{jk}:

$$\widehat{\psi}_{jk} = \frac{n_{000} n_{1jk}}{n_{100} n_{0jk}}.$$

To calculate the asymptotic variance of the estimator $\widehat{\psi}_{jk}$ (as n_1 and n_0 both approach infinity), we equivalently consider the asymptotic variance of $\log \widehat{\psi}_{jk}$. For $i = 1, 0$, the multinomial distribution $N_i \sim \text{Multinomial}(n_i, p_i)$ yields the asymptotic distribution of $\widehat{p}_i = (\widehat{p}_{i00}, \widehat{p}_{i01}, \widehat{p}_{i10}, \widehat{p}_{i11})^{\mathrm{T}}$ (Serfling 1980, pp.108-109):

$$\sqrt{n_i}\,(\widehat{p}_i - p_i) \xrightarrow{d} N(0, \Sigma_i) \tag{4}$$

as $n_i \to \infty$, where

$$\Sigma_i = \begin{pmatrix} p_{i00}(1-p_{i00}) & -p_{i00}p_{i01} & -p_{i00}p_{i10} & -p_{i00}p_{i11} \\ -p_{i01}p_{i00} & p_{i01}(1-p_{i01}) & -p_{i01}p_{i10} & -p_{i01}p_{i11} \\ -p_{i10}p_{i00} & -p_{i10}p_{i01} & p_{i10}(1-p_{i10}) & -p_{i10}p_{i11} \\ -p_{i11}p_{i00} & -p_{i11}p_{i01} & -p_{i11}p_{i10} & p_{i11}(1-p_{i11}) \end{pmatrix}$$

with the constraints $\sum_{j,k} \widehat{p}_{ijk} = 1$ and $\sum_{j,k} p_{ijk} = 1$ imposed. The asymptotic variances of the estimators $\widehat{\psi}_{jk}$ and $\widehat{\psi}$, or their logarithms, can be obtained using the delta method. Specifically, estimates of the asymptotic variances are

$$\widehat{\text{Avar}}\left(\log \widehat{\psi}_{jk}\right) = \frac{1}{n_{1jk}} + \frac{1}{n_{0jk}} + \frac{1}{n_{100}} + \frac{1}{n_{000}}$$

for $j, k = 0, 1$ and

$$\widehat{\text{Avar}}\left(\log \widehat{\psi}\right) = \sum_{i=0}^{1}\sum_{j=0}^{1}\sum_{k=0}^{1} \frac{1}{n_{ijk}}. \tag{5}$$

3 Interacting covariates with misclassification

In the presence of misclassification of the binary covariates, let X_a^* and X_s^* be the observed values of X_a and X_s, respectively. Let

$$\pi_{ia1} = P\left(X_a^* = 1 | X_a = 1, Y = i\right) \text{ and } \pi_{ia0} = P\left(X_a^* = 0 | X_a = 0, Y = i\right)$$

be respectively the *sensitivity* and *specificity* of X_a for the subpopulation with $Y = i$, and

$$\pi_{is1} = P\left(X_s^* = 1 | X_s = 1, Y = i\right) \text{ and } \pi_{is0} = P\left(X_s^* = 0 | X_s = 0, Y = i\right)$$

be respectively the *sensitivity* and *specificity* of X_s for the subpopulation with $Y = i$. Define

$$\Pi_{ia} = \begin{pmatrix} \pi_{ia0} & 1-\pi_{ia1} \\ 1-\pi_{ia0} & \pi_{ia1} \end{pmatrix} \text{ and } \Pi_{is} = \begin{pmatrix} \pi_{is0} & 1-\pi_{is0} \\ 1-\pi_{is1} & \pi_{is1} \end{pmatrix}.$$

For $i, j, k = 0, 1$, let

$$p^*_{ijk} = P\left(X^*_a = j, X^*_s = k | Y = i\right)$$

be the probabilities for the observed covariate measurements corresponding to the case or control subpopulation. Write $p^*_i = \left(p^*_{i00}, p^*_{i10}, p^*_{i01}, p^*_{i11}\right)$ for $i = 0, 1$.

We assume that

$$\begin{aligned} & P\left(X^*_a = j, X^*_s = k | X_a, X_s, Y\right) \\ = \; & P\left(X^*_a = j | X_a, X_s, Y\right) P\left(X^*_s = k | X_a, X_s, Y\right), \end{aligned}$$

$$P\left(X^*_a = j | X_a, X_s, Y\right) = P\left(X^*_a = j | X_a, Y\right),$$

and

$$P\left(X^*_s = j | X_a, X_s, Y\right) = P\left(X^*_s = j | X_s, Y\right).$$

The first assumption says that the observed measurements X^*_a and X^*_a are conditionally independent, given the true values X_a and X_s and the disease status. The second and third conditions require that the misclassification probability of one variable does not depend on the true value of the other variable, given the true value of the variable itself and the disease status. Under these assumptions, we express the probabilities p^*_{ijk} using the true probabilities p_{ijk}:

$$\begin{pmatrix} p^*_{i00} & p^*_{i01} \\ p^*_{i10} & p^*_{i11} \end{pmatrix} = \Pi_{ia} \begin{pmatrix} p_{i00} & p_{i01} \\ p_{i10} & p_{i11} \end{pmatrix} \Pi_{is}. \tag{6}$$

The identity (6) allows us to estimate the probability p_{ijk} using the estimates of p^*_{ijk} which can be obtained from the observed counts (Barron 1977). Let n^*_{ijk} represent the number of cases or controls with the observed measurement $\left(X^*_a = j, X^*_s = k\right)$ for $i, j, k = 0, 1$, as displayed in Table 1.

Using the same reasoning as for (3), we obtain the likelihood based on the observed data

$$L_{obs} = \prod_{i=0}^{1} \prod_{j=0}^{1} \prod_{k=0}^{1} \left(p^*_{ijk}\right)^{n^*_{ijk}}. \tag{7}$$

Maximizing the likelihood (7) with respect to the cell probabilities p^*_{ijk}, under the constraint $\sum_{j=0}^{1} \sum_{k=0}^{1} p^*_{ijk} = 1$ for $i = 0, 1$, gives their estimators

$$\widehat{p}^*_{ijk} = \frac{n^*_{ijk}}{n_i} \quad \text{for } i, j, k = 0, 1.$$

Applying (6), we obtain the estimators for the true cell probabilities p_{ijk}:

$$\begin{pmatrix} \widehat{p}_{i00} & \widehat{p}_{i01} \\ \widehat{p}_{i10} & \widehat{p}_{i11} \end{pmatrix} = \Pi^{-1}_{ia} \begin{pmatrix} \widehat{p}^*_{i00} & \widehat{p}^*_{i01} \\ \widehat{p}^*_{i10} & \widehat{p}^*_{i11} \end{pmatrix} \Pi^{-1}_{is}, \tag{8}$$

where the matrices Π_{ia} and Π_{is} are assumed invertible.

Table 1 Observed counts for case-control data

	$X^*_s = 0$		$X^*_s = 1$		
	$X^*_a = 0$	$X^*_a = 1$	$X^*_a = 0$	$X^*_a = 1$	Total
Case ($Y = 1$)	n^*_{100}	n^*_{110}	n^*_{101}	n^*_{111}	n_1
Control ($Y = 0$)	n^*_{000}	n^*_{010}	n^*_{001}	n^*_{011}	n_0

To describe the asymptotic variance $\widehat{p}_{ijk}$, we apply the delta method to the asymptotic distribution of $\left(\widehat{p}^*_{i00}, \widehat{p}^*_{i01}, \widehat{p}^*_{i10}, \widehat{p}^*_{i11}\right)^{\mathrm{T}}$ in combination of (8), where the asymptotic distribution of $\left(\widehat{p}^*_{i00}, \widehat{p}^*_{i01}, \widehat{p}^*_{i10}, \widehat{p}^*_{i11}\right)^{\mathrm{T}}$ is of the same form as (4) except for replacing p_{ijk} and $\widehat{p}_{ijk}$ with p^*_{ijk} and $\widehat{p}^*_{ijk}$ respectively, $i, j, k = 0, 1$.

4 Inference method with replicates

The foregoing method assumes that the misclassification probabilities are known, and it is useful for conducting sensitivity analyses where one may specify a class of plausible values of the sensitivity and specificity to evaluate the misclassification effects on estimation of quantities such as odds ratios ψ_{jk} or cell probabilities p_{ijk}.

In practice, misclassification probabilities are usually unavailable and must be estimated from additional data sources. Here we consider a situation where an independent sample with two repeated covariate measurements is available. In addition to the main study data displayed in Table 1, a second independent sample is available as shown in Table 2. As there is no information on the disease status for this independent sample, we assume the nondifferential misclassification mechanism in order to estimate the sensitivities and specificities. Namely, we assume that

$$\pi_{iaj} = \pi_{aj} \text{ and } \pi_{isj} = \pi_{sj},$$

where π_{aj} and π_{sj} are constants, $i, j = 0, 1$. Although no gold standard measurements of X_a and X_s via a validation subsample are available for this circumstance, the discrepancy between the two repeated measurements allows us to estimate the misclassification probabilities under certain assumptions.

To see this, we consider an estimation method of the π_{aj} and π_{sj} which separately uses the repeated measurements of X_a and X_s. We describe only estimation of the π_{aj} here; estimation of the π_{sj} is similar.

Let X^*_{a1} and X^*_{a2} denote the first and second observed measurements for X_a, respectively. Define

$$a_{ajk} = P\left(X^*_{a1} = j, X^*_{a2} = k\right) \text{ for } j, k = 0, 1$$

and $\alpha_a = P(X_a = 1)$. If assuming conditional independence between the first and second observed measurements for X_a:

$$\begin{aligned} & P\left(X^*_{a1} = j, X^*_{a2} = k | X_a = l\right) \\ = \; & P\left(X^*_{a1} = j | X_a = l\right) P\left(X^*_{a2} = k | X_a = l\right) \text{ for } j, k, l = 0, 1, \end{aligned}$$

Table 2 Two replicates of surrogate covariate measurements

First assessment	Second assessment			First assessment	Second assessment		
	$X^*_{a2} = 1$	$X^*_{a2} = 0$	Total		$X^*_{s2} = 1$	$X^*_{s2} = 0$	Total
$X^*_{a1} = 1$	n^*_{a11}	n^*_{a10}	n^*_{a1+}	$X^*_{s1} = 1$	n^*_{s11}	n^*_{s10}	n^*_{s1+}
$X^*_{a1} = 0$	n^*_{a01}	n^*_{a00}	n^*_{a0+}	$X^*_{s1} = 0$	n^*_{s01}	n^*_{s00}	n^*_{s0+}
Total			n^*_a				n^*_s

then we obtain $a_{a10} = a_{a01}$, and

$$\begin{aligned} a_{a11} &= \pi_{a1}^2\alpha_a + (1-\pi_{a0})^2(1-\alpha_a); \\ a_{a10} &= \pi_{a1}(1-\pi_{a1})\alpha_a + \pi_{a0}(1-\pi_{a0})(1-\alpha_a); \\ a_{a00} &= (1-\pi_{a1})^2\alpha_a + \pi_{a0}^2(1-\alpha_a). \end{aligned} \tag{9}$$

In (9), one equation is determined by the other two, implying that model parameters are unidentifiable unless additional assumptions are imposed. To consider a reduced parameter space, we take the prevalence α_a as given.

Let N_{ajk}^* be the number of pairs $\left(X_{a1}^* = j, X_{a2}^* = k\right)$ for $j,k = 0,1$. Then we have a multinomial distribution

$$\left(N_{a11}^*, N_{a10}^*, N_{a01}^*, N_{a00}^*\right) \sim \text{Multinomial}\left(n_a^*, a_{a11}, a_{a10}, a_{a01}, a_{a00}\right),$$

resulting in the likelihood,

$$L(\pi_{a1}, \pi_{a0}) = a_{a11}^{n_{a11}^*} \cdot a_{a10}^{n_{a10}^* + n_{a01}^*} \cdot a_{a00}^{n_{a00}^*}$$

where the constant is omitted, n_a^* is the number of paired assessments, a_{a11}, a_{a10} and a_{a00} are determined by (9) and constrained by $a_{a11} + 2a_{a10} + a_{a00} = 1$.

Estimation of parameters is carried out with the maximization of the likelihood $L(\pi_{a1}, \pi_{a0})$. The associated variance estimates are obtained from the observed information matrix, i.e., the negative of the second derivative matrix of $\log L(\pi_{a1}, \pi_{a0})$ evaluated at the estimates of parameters. To get rid of the constraints that the probabilities are bounded by 0 and 1, we reparameterize π_{a0} and π_{a1} by using the logit transformation when maximizing the likelihood.

To obtain estimates of regression parameters in model (1), we use the relationship (8) and obtain that

$$\widehat{p}_{ijk} = \frac{\widetilde{p}_{ijk}}{\left(\widehat{\pi}_{a0} + \widehat{\pi}_{a1} - 1\right)\left(\widehat{\pi}_{s0} + \widehat{\pi}_{s1} - 1\right)}$$

for $j,k = 0,1$, where

$$\begin{aligned} \widetilde{p}_{i00} &= \widehat{\pi}_{a1}\widehat{\pi}_{s1}\widehat{p}_{i00}^* - (1-\widehat{\pi}_{a1})\,\widehat{\pi}_{s1}\widehat{p}_{i10}^* - \widehat{\pi}_{a1}(1-\widehat{\pi}_{s1})\widehat{p}_{i01}^* + (1-\widehat{\pi}_{a1})(1-\widehat{\pi}_{s1})\widehat{p}_{i11}^*; \\ \widetilde{p}_{i10} &= -(1-\widehat{\pi}_{a0})\widehat{\pi}_{s1}\widehat{p}_{i00}^* + \widehat{\pi}_{a0}\widehat{\pi}_{s1}\widehat{p}_{i10}^* + (1-\widehat{\pi}_{a0})(1-\widehat{\pi}_{s1})\widehat{p}_{i01}^* - \widehat{\pi}_{a0}(1-\widehat{\pi}_{s1})\widehat{p}_{i11}^*; \\ \widetilde{p}_{i01} &= -\widehat{\pi}_{a1}(1-\widehat{\pi}_{s0})\widehat{p}_{i00}^* + (1-\widehat{\pi}_{a1})(1-\widehat{\pi}_{s0})\widehat{p}_{i10}^* + \widehat{\pi}_{a1}\widehat{\pi}_{s0}\widehat{p}_{i01}^* - (1-\widehat{\pi}_{a1})\widehat{\pi}_{s0}\widehat{p}_{i11}^*; \\ \widetilde{p}_{i11} &= (1-\widehat{\pi}_{a0})(1-\widehat{\pi}_{s0})\widehat{p}_{i00}^* - \widehat{\pi}_{a0}(1-\widehat{\pi}_{s0})\widehat{p}_{i10}^* - (1-\widehat{\pi}_{a0})\widehat{\pi}_{s0}\widehat{p}_{i01}^* + \widehat{\pi}_{a0}\widehat{\pi}_{s0}\widehat{p}_{i11}^*. \end{aligned} \tag{10}$$

Consequently, the log odds ratios are estimated by

$$\log\left(\widehat{\psi}_{jk}\right) = \log\left(\frac{\widehat{p}_{000}\cdot\widehat{p}_{1jk}}{\widehat{p}_{100}\cdot\widehat{p}_{0jk}}\right) = \log\left(\frac{\widetilde{p}_{000}\cdot\widetilde{p}_{1jk}}{\widetilde{p}_{100}\cdot\widetilde{p}_{0jk}}\right) \tag{11}$$

for $(j,k) \neq (0,0)$. The variance of $\log\left(\widehat{\psi}_{jk}\right)$ can be obtained by applying the delta method to the variance of $\left(\widehat{\pi}_{a1}, \widehat{\pi}_{a0}, \widehat{\pi}_{s1}, \widehat{\pi}_{s0}, \widehat{p}_0^{*\mathrm{T}}, \widehat{p}_1^{*\mathrm{T}}\right)^{\mathrm{T}}$ which is a diagonal block matrix with variances of $(\widehat{\pi}_{a1}, \widehat{\pi}_{a0})^{\mathrm{T}}$, $(\widehat{\pi}_{s1}, \widehat{\pi}_{s0})^{\mathrm{T}}$, $\widehat{p}_0^*$ and $\widehat{p}_1^*$ being the diagonal blocks. The derivatives of $\log\left(\widehat{\psi}_{jk}\right) = \log(\widetilde{p}_{000}) + \log(\widetilde{p}_{1jk}) - \log(\widetilde{p}_{100}) - \log(\widetilde{p}_{0jk})$ with respect to the parameters can be easily obtained via (10).

Finally, as noted by a referee, certain constraints underlie the estimates of log odds ratios (11), which are reflected by the positivity of the probabilities in (10). These constraints essentially require the misclassification probabilities to be upper bounded properly to ensure that the observed surrogate measurements are relevant and useful. In other words,

misclassification effects can only be addressed when they are not arbitrarily substantial, and this makes intuitive sense. For instance, when a misclassification probability, say $P\left(X_a^* = 0|X_a = 1\right)$, is bigger than $1/2$, then the observed measurements X_a^* carry useless information of X_a; using such observations to estimate the model parameter, no matter how an estimation method is developed, is even worse than using artificial data generated from flipping a fair coin.

5 Numerical analysis

In this section, we analyze the motivating example to illustrate the usage of the proposed method and conduct numerical studies to assess the performance of our method.

5.1 Data analysis

We analyze the case-control data discussed by Duffy et al. (1989) and described in Section 1. For any subject, let X_a be a binary variable indicating whether or not the alcohol consumption is more than 9.3 g ethanol/day, and let X_s be a binary variable indicating whether or not the lifetime cigarette-years of the subject is more than 300. Table 3 records the data of the main study, where one breast cancer case has missing observations and we ignore this in the analysis. In addition, there was an independent study available on 100 women who were neither cases nor controls. Repeated measurements of X_a and X_s were collected for those women on two occasions, and the measurements are given in Table 4 where one subject has missing observations of X_s.

We analyze the data using the proposed method described in Section 4 and the naive method with misclassification in X_a and X_s ignored, called Analysis 1 and Analysis 2, respectively. To remove the constraints of the specificities and sensitivities, we consider the reparameterization:

$$\widetilde{\pi} = \frac{\exp(\delta)}{1 + \exp(\delta)},$$

where $\widetilde{\pi}$ represents $\pi_{a1}, \pi_{a0}, \pi_{s1}$ or π_{s0}, and δ is the corresponding parameter which takes a value in $(-\infty, +\infty)$.

As comparisons, a referee suggested to further conduct two analyses, called Analysis 3 and Analysis 4. In Analysis 3, we take the misclassification probabilities as known and let their values be determined by the estimated specificities and sensitivities obtained from Analysis 1. In Analysis 4, we pretend the second sample is a validation sample where the measurements from the first assessment were taken as the true values and the measurements from the second assessment were regarded as surrogate measurements; sensitivities and specificities are then estimated from the relative frequencies using this artificial validation sample.

Table 3 Breast cancer case-control study: main study data

	$X_s^* = 0$		$X_s^* = 1$		
	$X_a^* = 0$	$X_a^* = 1$	$X_a^* = 0$	$X_a^* = 1$	Total
$Y = 1$	268	82	61	39	450
$Y = 0$	305	70	56	20	451
Total	573	152	117	59	901

Table 4 Breast Cancer Case-Control Study: Replicates of Surrogate Measurements

First assessment	Second assessment			First assessment	Second assessment		
	$X_a^* = 1$	$X_a^* = 0$	Total		$X_s^* = 1$	$X_s^* = 0$	Total
$X_a^* = 1$	18	6	24	$X_s^* = 1$	11	2	13
$X_a^* = 0$	7	69	76	$X_s^* = 0$	2	84	86
Total			100				99

The analysis results are reported in Table 5, where EST, SEM and 95% CI represent estimates, model-based standard errors and 95% confidence intervals for the parameters, respectively. Relative to those produced by the method with misclassification incorporated (Analysis 1), the naive analysis (Analysis 2) yields attenuated point estimates for β_a and β_{as}, and leads to an inflated estimate of β_s. Analysis 1 produces larger standard errors than Analysis 2, which is consistent with the typical patterns observed in the analysis with measurement error models in the literature. The comparison between the results of Analyses 1 and 3 confirms the theoretical property that Analysis 3 produces the same point estimates of the response parameters as Analysis 1 does, but it yields smaller variance estimates than those produced from Analysis 1. While it is not possible to directly compare Analysis 4 to Analysis 1 or Analysis 3, the comparison of Analysis 4 to Analysis 2 reveals the same pattern as the comparison between Analyses 1 and 2. All the analyses suggest that none of smoking, alcohol consumption, and their interaction are statistically significant.

5.2 Sensitivity analysis

In the previous subsection the misclassification probabilities are estimated based on a small set of replicated surrogate measurements, whose accuracy may be questionable due to the small size of the data. We now investigate the effect of misclassification of the alcohol and smoking factors on the estimation of the odds ratios when misclassification probabilities are set differently. Three scenarios are considered: there is misclassification on alcohol factor only, on smoking factor only, and on both the alcohol and smoking factors. The sensitivity and specificity are employed to specify the (correct) classification rates; setting these quantities to be 1 corresponds to the case without misclassification. To ensure the nonnegativity of the probabilities p^*, specification of the sensitivity and specificity is subject to underlying constraints, as discussed in Section 4.

Table 5 Analysis results for the breast cancer case-control study

	Analysis 1			Analysis 2		
	EST	SEM	95% CI	EST	SEM	95% CI
β_a	0.340	0.307	(-0.261, 0.941)	0.288	0.183	(-0.071, 0.646)
β_s	0.175	0.293	(-0.400, 0.750)	0.215	0.203	(-0.183, 0.613)
β_{as}	0.372	0.520	(-0.647, 1.391)	0.295	0.379	(-0.447, 1.037)
	Analysis 3			Analysis 4		
	EST	SEM	95% CI	EST	SEM	95% CI
β_a	0.340	0.224	(-0.099, 0.778)	0.454	0.349	(-0.229, 1.138)
β_s	0.175	0.259	(-0.334, 0.683)	0.153	0.291	(-0.418, 0.724)
β_{as}	0.372	0.497	(-0.603, 1.346)	0.410	0.686	(-0.934, 1.754)

Noting that estimates of the log odds ratios are determined by (11) which includes the terms $\widetilde{p}_{ijk}$ for $i, j, k = 0, 1$ and that, by (10), $\widetilde{p}_{ijk}$ depends on the sensitivity and the specificity in the same manner, one might expect that the change of a log odds ratio relative to the sensitivity would behave in the same manner as that to the specificity. However, this speculation is not necessarily visualizable because the *magnitude* of the change can be different due to the dependence of the log odds ratios on estimated probabilities $\widehat{p}^{*}_{ijk}$ obtained from the observed data. In other words, *visual* effects of the sensitivity and the specificity on changes of log odds ratios can be noticeably different, which is driven by the *actual observed* data in Table 1. This is reflected in our sensitivity analyses here.

Figure 1 shows the change of the log odds ratios according to the change of the sensitivity and specificity for the alcohol factor while keeping the sensitivity and specificity for the smoking factor to be 1. With a given specificity (i.e., π_{a0}) or sensitivity (i.e., π_{a1}), the log odds ratio $\log(\psi_{10})$ tends to decrease as the sensitivity (i.e., π_{a1}) or specificity (i.e., π_{a0}) of alcohol factor increases; but the change rates for them are not the same. On the other hand, the log odds ratio $\log(\psi_{01})$ increases as the sensitivity (i.e., π_{a1}) of alcohol increases when the specificity (i.e., π_{a0}) is kept fixed; the log odds ratio $\log(\psi_{01})$ appears less sensitive to the change of the specificity (i.e., π_{a0}) when the sensitivity (i.e., π_{a1}) is given. Regarding the log odds ratio $\log \psi$, we notice that its value is affected the change in the sensitivity and specificity of the alcohol.

Figure 2 presents the changes of the log odds ratios according to the change of the sensitivity and specificity for the smoking factor while keeping the sensitivity and specificity of the alcohol factor specified as 1. The log odds ratios $\log(\psi_{10})$ appears to increase as the sensitivity (i.e., π_{s1}) of the smoking factor increases with the specificity (i.e., π_{s0}) kept fixed; whereas when the the sensitivity (i.e., π_{s1}) of the smoking factor is fixed, the log odds ratios $\log(\psi_{10})$ tends to decrease as the specificity (i.e., π_{s0}) of the smoking factor increases; Again, the change in sensitivity and specificity of the smoking factor affects the value of $\log(\psi)$.

Figure 3 shows how the log odds ratios may change relative to the change of the sensitivity and specificity for both the alcohol and smoking factors. While any circumstances can be considered, here we confine our attention to the scenario where the sensitivity for the alcohol and smoking factors is equal and the specificity for these two factors is common. It is evident that the values of both the sensitivity and specificity of the alcohol and smoking factors have the impact on estimation of the log odds ratios while the magnitudes can be different from case to case.

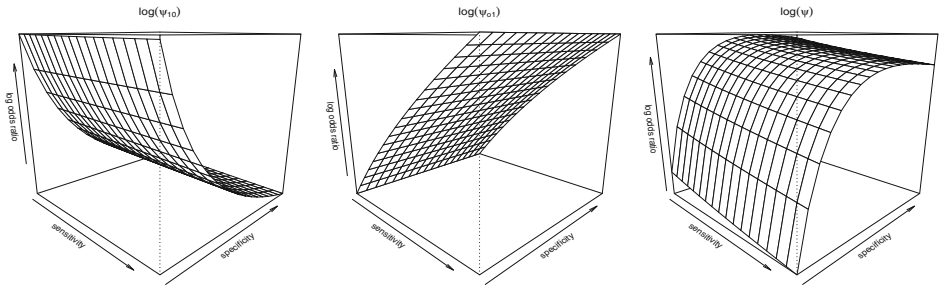

Fig. 1 Sensitivity analyses of the log odds ratios to the changes of π_{a0} and π_{a1} where π_{s0} and π_{s1} are set as 1. The *left plot* is for $\log \psi_{10}$, the *center plot* is for $\log \psi_{01}$, and the *right plot* is for $\log \psi$

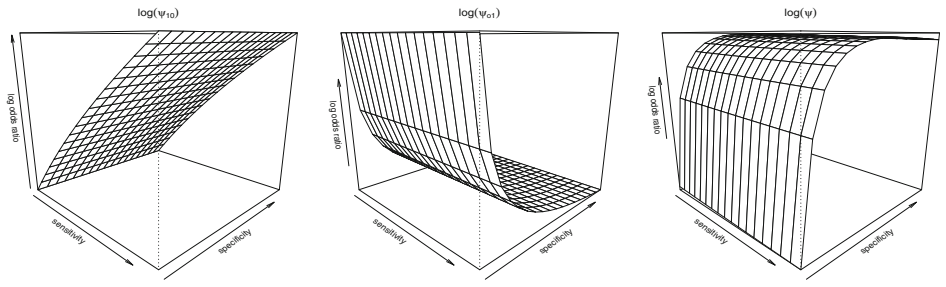

Fig. 2 Sensitivity analyses of the log odds ratios to the changes of π_{s0} and π_{s1} where π_{a0} and π_{a1} are set as 1. The left plot is for log ψ_{10}, the center plot is for log ψ_{01}, and the right plot is for log ψ

5.3 Simulation study

In this subsection, we conduct simulation studies to assess the performance of the proposed method and to demonstrate the impact of ignoring the misclassification in the analysis.

We consider a setting similar to one of Zhang et al. (2008). Let X_a be generated from a binomial distribution BIN(1,0.5) and let X_s be generated from a binomial distribution BIN(1,0.5). Response Y is generated from model (1) where we set $\beta_a = \beta_s = \beta_{as} = \log(2.0)$, and $\beta_0 = -3.0$. For the sensitivity and specificity, we consider two settings: (I) $\pi_{a0} = \pi_{a1} = 0.8, \pi_{s0} = \pi_{s1} = 0.9$, and (II) $\pi_{a0} = \pi_{a1} = 0.9, \pi_{s0} = \pi_{s1} = 0.95$.

First, we generate a large number of individuals, say, 200000 individuals, which are treated as the underlying population. Then we randomly select n_1 cases and n_0 controls from this population to form a main study sample. We consider three scenarios with different sizes of cases and controls. In the first scenario, we take $n_1 = n_0 = 1000$; in the second scenario, we take $n_1 = n_0 = 500$; and in the third scenario, we take $n_1 = 200$ and $n_0 = 600$. To generate a second sample of replicates, we randomly select n^* individuals from the underlying population so that each individual has two repeated surrogate measurements for each of X_a and X_s. We consider two scenarios, called Scenario R_1 and Scenario R_2, where n^* is set as 100 and 500, respectively.

For each parameter configuration, we simulate 500 data sets and analyze the data using both the the proposed method and the naive method which disregards the misclassification feature. We report the bias (Bias), the model-based standard error (SEM), and the 95% confidence interval coverage rate (CR%), and the results are reported in Tables 6, 7 and 8, each corresponding to a size scenario.

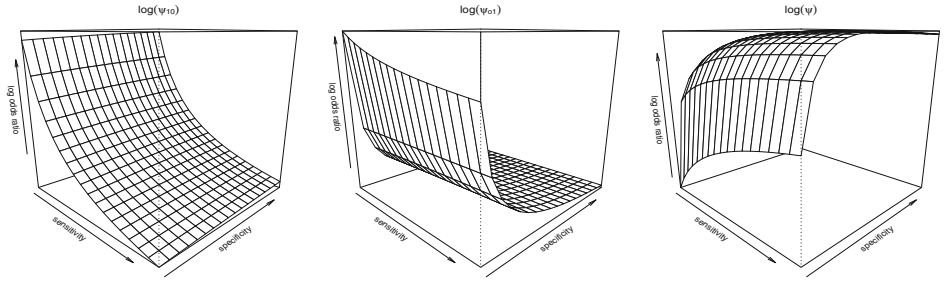

Fig. 3 Sensitivity analyses of the log odds ratios to the changes of π_0 and π_1, where π_{s0} and π_{a0} are set to be identical, and π_{s1} and π_{a1} are set to be identical. The *left plot* is for log ψ_{10}, the *center plot* is for log ψ_{01}, and the *right plot* is for log ψ

Table 6 Simulation results for the main study data with 1000 cases and 1000 controls

			β_a			β_s			β_{as}		
	Setting	Method	Bias	SEM	CR%	Bias	SEM	CR%	Bias	SEM	CR%
R_1	I	Naive	-0.211	0.139	68.0	0.039	0.142	93.6	-0.413	0.184	43.6
		Proposed	0.054	0.438	99.2	-0.018	0.459	98.2	0.104	0.604	98.4
	II	Naive	-0.079	0.145	92.0	0.046	0.140	94.4	-0.249	0.193	75.6
		Proposed	0.042	0.256	98.0	-0.011	0.264	96.2	0.034	0.349	97.8
R_2	I	Naive	-0.217	0.144	67.8	0.034	0.143	94.2	-0.411	0.191	39.4
		Proposed	0.015	0.350	95.8	0.026	0.328	96.6	-0.003	0.475	96.6
	II	Naive	-0.087	0.142	93.2	0.034	0.152	95.0	-0.236	0.197	77.4
		Proposed	0.011	0.212	97.0	-0.005	0.223	96.0	0.008	0.301	94.8

It is clear that ignoring misclassification yields biased estimates of the parameters, and the coverage rates for the 95% confidence intervals considerably deviate from the nominal level. On the contrary, the proposed method yields much improved estimation results with a lot smaller biases. As a trade-off of improving point estimation, variances of the proposed estimators are bigger than those of the naive estimators, which has also been observed in other problems concerning measurement error or misclassification (e.g., Carroll et al. 2006; Yi 2017). However, jointly reporting point estimation and associated variability, the proposed method produces much better coverage rates for the 95% confidence intervals. More specifically, with a given scenario of R_1 or R_2, the proposed method tends to produce better results for Setting I than for Setting II, as expected. With a given setting of I or II, standard errors obtained from Scenario R_2 are smaller than those obtained from Scenario R_1.

6 Discussion and extensions

Dichotomized covariates are very common in medical studies and misclassification of these covariates happens frequently in the data collection process. It is important to incorporate such a feature in the data analysis; otherwise, biased results are usually derived. In this article, we investigate misclassification effects of error-prone binary covariates on the estimation of risk measures for case-control studies and develop a valid inference method for addressing misclassification effects. Our development is carried out under a practical setting where a validation sample is impossible but repeated measurements of

Table 7 Simulation results for the main study data with 500 cases and 500 controls

			β_a			β_s			β_{as}		
	Setting	Method	Bias	SEM	CR%	Bias	SEM	CR%	Bias	SEM	CR%
R_1	I	Naive	-0.218	0.205	82.4	0.017	0.208	94.8	-0.401	0.272	67.2
		Proposed	0.045	0.618	99.4	-0.034	0.657	98.8	0.125	0.874	98.8
	II	Naive	-0.093	0.214	92.4	0.046	0.199	96.0	-0.248	0.270	86.0
		Proposed	0.010	0.336	96.2	-0.003	0.346	97.8	0.029	0.457	97.2
R_2	I	Naive	-0.210	0.204	83.2	0.024	0.204	94.0	-0.399	0.274	71.0
		Proposed	0.033	0.506	97.6	0.004	0.496	96.8	0.042	0.709	95.4
	II	Naive	-0.090	0.204	94.2	0.029	0.201	95.6	-0.224	0.256	88.8
		Proposed	0.012	0.303	97.4	-0.013	0.294	97.0	0.028	0.394	97.0

Table 8 Simulation results for the main study data with 200 cases and 600 controls

			β_a			β_s			β_{as}		
	Setting	Method	Bias	SEM	CR%	Bias	SEM	CR%	Bias	SEM	CR%
R_1	I	Naive	-0.216	0.273	89.6	0.037	0.258	96.4	-0.405	0.346	79.4
		Proposed	0.069	0.801	99.6	0.017	0.757	100.0	0.093	1.044	99.0
	II	Naive	-0.099	0.306	93.0	0.034	0.301	94.0	-0.231	0.365	92.0
		Proposed	0.026	0.517	98.2	-0.022	0.562	97.2	0.053	0.671	98.2
R_2	I	Naive	-0.220	0.282	86.6	0.039	0.281	95.6	-0.404	0.362	78.6
		Proposed	0.056	0.760	98.4	0.037	0.791	99.2	0.022	1.047	99.0
	II	Naive	0.067	0.289	96.0	0.063	0.296	96.2	-0.257	0.371	88.8
		Proposed	0.052	0.452	98.4	0.043	0.461	98.2	-0.021	0.597	96.8

error-contaminated variables are available. Numerical studies demonstrate satisfactory performance of our method.

Our method is motivated by the breast cancer case-control data discussed by Duffy et al. (1989) which contain two error-prone binary covariates and each has two replicates of surrogate measurements. It is possible to extend our method to more general settings where error-prone binary covariates may be more than two, or/and replicates of surrogate measurements can be arbitrary, or/and error-free covariates are also present. Here we outline three extensions.

6.1 Extension 1: replicates are more than 2

If there are m repeated measurements of each of covariate in model (1), then the development in Section 4 can be generalized as follows with the discussion on one of the two covariates.

Let X^*_{aj} denote the jth observed measurement of X_a for $j = 1, \ldots, m$ where m is an integer greater than 2. Let $\alpha_a = P(X_a = 1)$ be the prevalence which is assumed known. Define

$$a_{aj_1 \ldots j_m} = P\left(X^*_{a1} = j_1, \ldots, X^*_{am} = j_m\right) \quad \text{for } j_k = 0, 1 \text{ and } k = 1, \ldots, m.$$

Without loss of generality, we assume that these m replicates are independently collected, thus yielding

$$\begin{aligned} a_{aj_1 \ldots j_m} &= P\left(X^*_{a1} = j_1, \ldots, X^*_{am} = j_m, X_a = 1\right) + P\left(X^*_{a1} = j_1, \ldots, X^*_{am} = j_m, X_a = 0\right) \\ &= \prod_{k=1}^{m} P\left(X^*_{ak} = j_k | X_a = 1\right) P\left(X_a = 1\right) + \prod_{k=1}^{m} P\left(X^*_{ak} = j_k | X_a = 0\right) P(X_a = 0) \\ &= \prod_{k=1}^{m} \pi_{a1}^{j_k} (1 - \pi_{a1})^{1 - j_k} \alpha_a + \prod_{k=1}^{m} (1 - \pi_{a0})^{j_k} \pi_{a0}^{1 - j_k} (1 - \alpha_a). \end{aligned}$$

Suppose there are n^*_a measurements in total for covariate X_a. Let $N^*_{aj_1 \ldots j_m}$ be the number of outcome $\left(X^*_{a1} = j_1, \ldots, X^*_{am} = j_m\right)$ for $j_k = 0, 1$ and $k = 1, \ldots, m$, and let $N^*_a = \left(N^*_{aj_1 \ldots j_m} : j_k = 0, 1; k = 1, \ldots, m\right)^{\mathrm{T}}$. Then we have a multinomial distribution

$$N^*_a \sim \text{Multinomial}\left(n^*_a, a_a\right),$$

where $a_a = \left(a_{aj_1\ldots j_m} : j_k = 0, 1; k = 1, \ldots, m\right)^{\mathrm{T}}$ with $\sum_{j_k=0,1;k=1,\ldots,m} a_{aj_1\ldots j_m} = 1$, resulting in the likelihood

$$L(\pi_{a1}, \pi_{a0}) = \prod_{j_k=0,1;k=1,\ldots,m} \left(a_{aj_1\ldots j_m}\right)^{n^*_{aj_1\ldots jm}}$$

where the constant is omitted. Estimation of parameters π_{a1} and π_{a0} is carried out with the maximization of the likelihood $L(\pi_{a1}, \pi_{a0})$, and the associated variance estimates are obtained from the negative of the second derivative matrix of $\log L(\pi_{a1}, \pi_{a0})$ evaluated at the estimates of parameters. Other development in the preceding sections can then carry through.

6.2 Extension 2: covariates are more than 2

When the number of binary covariates is greater than 2, model (1) can be generalized. To be specific, let X_l denote the jth binary covariate for $j = 1, \ldots, p$ where p is an integer greater than 2. Then model (1) can be generalized as

$$\log\left\{\frac{P\left(Y = 1|X_1, \ldots, X_p\right)}{P\left(Y = 0|X_1, \ldots, X_p\right)}\right\} = \beta_0 + \sum_{l=1}^{p} \beta_l X_l + \sum_{j<k} \beta_{jk} X_j X_k, \tag{12}$$

where the β_0, β_l and β_{jk} are the regression parameters for $l = 1, \ldots, p$ and $1 \le j < k \le p$.

These parameters can be interpreted in terms of the odds ratios defined for the retrospective sampling framework. Specifically, let

$$q_i = P\left(X_1 = 0, \ldots, X_p = 0|Y = i\right)$$

and

$$p_{i(k)} = P\left(X_1 = 0, \ldots, X_{k-1} = 0, X_k = 1, X_{k+1} = 0, \ldots, X_p = 0|Y = i\right)$$

for $i = 0$ or 1 and $k = 1, \ldots, p$. Let ψ_k be the odds ratio for cases versus controls with $\left(X_1 = 0, \ldots, X_{k-1} = 0, X_k = 1, X_{k+1} = 0, \ldots, X_p = 0\right)$ compared with the baseline category $\left(X_1 = 0, \ldots, X_p = 0\right)$:

$$\psi_k = \frac{q_0 p_{1(k)}}{q_1 p_{0(k)}}$$

for $k = 1, \ldots, p$.

For $i = 0$ or 1 and $1 \le j < k \le p$, let

$$p_{i(jk)} = P\left(X_j = X_k = 1; X_l = 0 : l \ne j, l \ne k|Y = i\right).$$

Define

$$\psi_{jk} = \frac{q_0 p_{1(jk)}}{q_1 p_{0(jk)}} \text{ and } \phi_{jk} = \frac{\psi_{jk}}{\psi_j \psi_k}.$$

Then

$$\beta_l = \log \psi_l \text{ and } \beta_{jk} = \log \phi_{jk}$$

for $l = 1, \ldots, p$ and $1 \le j < k \le p$. Other development in the preceding sections can then carry through with a more complex exposition.

We note that model (12) reflects the main effects as well as all pairwise interactions among the covariates. The three-way or higher order interactions among the covariates are not included, which are virtually assumed to be zero. In problems for which such

interactions are of interest, one may modify model (12) by adding those terms with additional parameters introduced. In principle, any order of interactions among the covariates may be included in the model until a saturated model is formed. The interpretation of the associated parameters would be modified accordingly.

6.3 Extension 3: error-free covariates are also present

Model (1) can be modified to accommodate settings with error-prone covariates as well. Let Z denote the vector of error-free risk factors of a disease. The prospective logistic regression model is then written as

$$\log\left\{\frac{P(Y=1|X_a,X_s,Z)}{P(Y=0|X_a,X_s,Z)}\right\}=\beta_0+\beta_aX_a+\beta_sX_s+\beta_{as}X_aX_s+\beta_z^{\mathrm{T}}Z, \tag{13}$$

where $\beta_0,\beta_a,\beta_s,\beta_{as}$ and β_z are the regression parameters. The parameters β_0,β_a,β_s, and β_{as} can be interpreted in the same manner as (2) except that the associated conditional probabilities p_{ijk} need to be modified as

$$p_{ijk}=P\left(X_a=j,X_s=k|Y=i,Z\right)$$

with error-free covariates Z being controlled. Estimation of the model parameters may then be carried out using the likelihood method.

To conclude, we comment that the development of Section 4 is based on the assumption of the nondifferential misclassification mechanism. This assumption allows us to estimate the sensitivities and specificities using a separate sample from the main study which has repeated surrogate measurements of covariates only but not measurements of the disease status. Such an assumption, however, may be too restrictive for some applications, especially for retrospective studies. In such instances, conducting sensitivity analyses can be a viable way to allow us not to impose the nondifferential misclassification mechanism but enable us to explore the impact of misclassification on inference results. Finally, our work here focuses on estimation of the model parameters. It is also interesting to develop procedures for hypothesis testing to incorporate misclassification effects along the lines of Bross (1954).

Acknowledgements
The authors thank two anonymous referees whose comments improved the presentation of the manuscript. The research was supported by the Natural Sciences and Engineering Research Council of Canada.

Authors contribution
Both authors share contribution to the manuscript. GY proposed the research idea, and both authors together developed the methodology. WH worked on the computational analysis and GY drafted the manuscript. Both authors read and approved the final manuscript.

Competing interest
The authors declare that they have no competing interests.

Publisher's Note

Author details
[1] Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Canada. [2] Department of Statistical and Actuarial Sciences, University of Western Ontario, London, Canada.

Received: 5 April 2017 Accepted: 11 July 2017
Published online: 30 October 2017

References

Armstrong, BG, Whittemore, AS, Howe, GR: Analysis of case-control data with covariate measurement error: Application to diet and colon cancer. Stat. Med. **8**, 1151–1163 (1989)
Barron, BA: The Effects of misclassification on the estimation of relative risk. Biometrics. **33**, 414–418 (1977)
Breslow, NE, Cain, KC: Logistic regression for two-stage case-control data. Biometrika. **75**, 11–20 (1988)
Breslow, NE, Day, NE: Statistical Methods in Cancer Research, Volume I - The Analysis of Case-Control Studies. International Agency for Research on Cancer, Lyon (1980)
Bross, I: Misclassification in 2×2 tables. Biometrics. **10**, 478–486 (1954)
Carroll, RJ, Gail, MH, Lubin, JH: Case-control studies with errors in covariates. J. Am. Stat. Assoc. **88**, 185–199 (1993)
Carroll, RJ, Ruppert, D, Stefanski, LA, Crainiceanu, CM: Measurement Error in Nonlinear Models. 2nd ed. Chapman & Hall/CRC, Boca Raton (2006)
Carroll, RJ, Wang, S, Wang, CY: Prospective analysis of logistic case-control studeis. J. Am. Stat. Assoc. **90**, 157–169 (1995)
Chu, H, Cole, SR, Wei, Y, Ibrahim, JG: Estimation and inference for case-control studies with multiple nongold standard exposure assessments: with an occupational health application. Biostatistics. **10**, 591–602 (2009)
Duffy, SW, Rohan, TE, Day, NE: Misclassification in more than one factor in a case-control study: A combination of Mantel-Haenszel and maximum likelihood approaches. Stat. Med. **8**, 1529–1536 (1989)
Forbes, AB, Santner, TJ: Estimators of odds ratio regression parameters in matched case-control studies with covariate measurement errror. J. Am. Stat. Assoc. **90**, 1075–1084 (1995)
Gustafson, P, Le, ND, Saskin, R: Case-control snalysis with partial knowledge of exposure misclassification probabilities. Biometrics. **57**, 598–609 (2001)
Lyles, RH: A note on estimating crude odds ratios in case-control studies with differentially misclassified exposure. Biometrics. **58**, 1034–1036 (2002)
Mak, TSH, Best, N, Rushton, L: Robust Bayesian sensitivity analysis for case-control studies with uncertain exposure misclassification probabilities. Int. J. Biostat. **11**, 135–149 (2015)
Marinos, AT, Tzonou, AJ, Karantzas, ME: Experimental quantiles of epidemiological indices in case-control studies with non-differential misclassification. Stat. Med. **14**, 1291–1306 (1995)
Morrissey, M, Spiegelman, D: Matrix methods for estimating odds ratios with misclassified exposure data: Extensions and comparisons. Biometrics. **55**, 338–344 (1999)
Prentice, RL, Pyke, R: Logistic disease incidence models and case-control studies. Biometrika. **66**, 403–411 (1979)
Prescott, GJ, Garthwaite, PH: Bayesian analysis of misclassified binary data from a matched case-control study with a validation substdy. Stat. Med. **24**, 379–401 (2005)
Roeder, K, Carroll, RJ, Lindsay, BG: A semiparametric mixture approach to case-control studies with error in covariables. J. Am. Stat. Assoc. **91**, 722–732 (1996)
Schlesselman, JJ: Case-Control Studies: Design, Conduct, Analysis. Oxford University Press, Oxford (1982)
Serfling, RJ: Approximation Theorems of Mathematical Statistics. Wiley, New York (1980)
Tang, L, Lyles, RH, Ye, Y, Lo, Y, King, CC: Extended matrix and inverse matrix methods utilizing internal validation data when both disease and exposure status are misclassified. Epidemiol. Methods. **2**, 49–66 (2013)
Yi, GY: Statistical Analysis with Measurement Error or Misclassification: Strategy, Method and Application. Springer Science+Business Media LLC, New York (2017)
Zhang, L, Mukherjee, B, Ghosh, M, Gruber, S, Moreno, V: Accounting for error due to misclassification of exposures in case-control studies of gene-environment interaction. Stat. Med. **27**, 2756–2783 (2008)