Biostatistics (2019) **0**, 0, *pp.* 1–18
doi:10.1093/biostatistics/kxz012

Matched case–control data with a misclassified exposure: what can be done with instrumental variables?

CHRISTOPHER M. MANUEL, SAMIRAN SINHA*, SUOJIN WANG

Department of Statistics, Texas A&M University, College Station, TX 77843-3143, USA
sinha@stat.tamu.edu

Summary

Matched case–control studies are used for finding the association between a disease and an exposure after controlling the effect of important confounding variables. It is a known fact that the disease–exposure association parameter estimators are biased when the exposure is misclassified, and a matched case–control study is of no exception. Any bias correction method relies on validation data that contain the true exposure and the misclassified exposure value, and in turn the validation data help to estimate the misclassification probabilities. The question is what we can do when there are no validation data and no prior knowledge on the misclassification probabilities, but some instrumental variables are observed. To answer this unexplored and unanswered question, we propose two methods of reducing the exposure misclassification bias in the analysis of a matched case–control data when instrumental variables are measured for each subject of the study. The significance of these approaches is that the proposed methods are designed to work without any validation data that often are not available when the true exposure is impossible or too costly to measure. A simulation study explores different types of instrumental variable scenarios and investigates when the proposed methods work, and how much bias can be reduced. For the purpose of illustration, we apply the methods to a nested case–control data sampled from the 1989 US birth registry.

Keywords: Conditional likelihood; Confounding variables; Instrumental variables; Logistic model; Low birthweight data; Misclassification.

1. Introduction

Matched case–control designs are utilized to assess the association between an outcome and a treatment/exposure in observational studies. To control the effect of confounding variables, cases are usually matched with controls based on the potential confounding variables, and this forms a stratum or a matched set. A matched case–control data set contains many such matched sets. Typically, a logistic model is assumed for modeling the disease incidence in terms of the exposure and other covariates (prognostic factors), and the parameters are estimated by maximizing a conditional likelihood function. The estimators of the parameters are biased when the exposure variable is mismeasured or misclassified which may

*To whom correspondence should be addressed.

happen due to many reasons including but not limited to recall error and misreporting. To be specific, we shall not consider the case where a continuous exposure variable is measured with error.

To address this misclassification bias of a binary exposure variable in a matched study, several Bayesian and frequentist methods have been developed. Briefly, we review some key papers in this paradigm. For the ease of our following discussion, we shall use Y, X, and W to denote the binary disease indicator, the binary exposure variable, and the misclasssified version of X, respectively. Typically, in the main study X is not observed, only Y and W are observed, while in the validation data Y, X, and W are observed. Usually validation data have a much smaller size than the main study, and the validation data are used for estimating the misclassification probabilities. To remove the misclassification bias one has to take into account these misclassification probabilities in the analysis.

Rice (2003) proposed a full likelihood approach when a binary exposure is misclassified in matched case–control data. Particularly, he wrote the likelihood in a multinomial model format where the cell probabilities are functions of the common odds ratio and non-differential misclassification probabilities. In the presence of validation data, he then proposed to estimate the parameters in a Bayesian framework. Prescott and Garthwaite (2005) proposed a Bayesian method of estimating the parameters when a binary exposure variable is misclassified in matched case–control data, but correct values of the exposure are available on a number of matched sets. In one of the three methods that they proposed, the authors first estimated the parameter from the validation data where the true values of the exposure are available. Next they proposed to use the first stage analysis result as the prior distribution for the second stage analysis where the likelihood is based on the matched sets in which only the misclassified exposure values are available. Chu *and others* (2010) proposed a Bayesian method where the main parameters of interest were the exposure prevalence among cases and controls, and specificities and sensitivities among cases and controls. They then developed a Bayesian inference with flexible priors on each of these parameters. Liu *and others* (2009) considered a Bayesian solution to the misclassification bias in a 1:1 matched case–control study. In the case of non-availability of validation data, they proposed to use expert prior knowledge on the disease–exposure association and misclassification probabilities. Morrissey and Spiegelman (1999) compared the matrix method, the inverse matrix method, and the maximum likelihood approach under the differential and non-differential misclassification scenarios. Lyles *and others* (2007) considered two robust weighted estimators that used two validation data sets, an internal validation data set and an external and less precise validation data set, for estimating the odds ratio parameter in the presence of a misclassified binary exposure in a regular case–control study. In a slightly different approach, Duffy *and others* (2003) proposed to replace the terms in the Mantel–Haenszel estimator by the corresponding conditional expected values given the observed data. These conditional expected values were functions of the misclassification probabilities that were derived from the validation data. Due to the difficulty in computing the large sample variance formula, they provided a Bayesian credible interval for the common odds ratio parameter using non-informative priors.

The important difference between the previously cited articles and our problem is that in our set-up there are no validation data. That means true values of the binary exposure X are never observed in any part of the data. Instead, a set of instrumental variables are available for each study subject. According to the standard definition (Greenland, 2000), the instrumental variables are correlated with X, but uncorrelated with the response and confounding variables conditional on X. Our goal is to provide a statistical method of inference in this context. Although instrumental variables are commonly used to solve the problem of endogeneity, no one has ever used instrumental variables to reduce misclassification bias in a retrospective case–control or matched case–control study. Particularly, instrumental variables have been used for consistent estimation of parameters of the linear regression or polynomial regression model in the presence of covariate measurement error of a numeric variable (Bound and Krueger, 1991;

Hausman *and others*, 1991). However, the use of instrumental variables for handling misclassification bias is limited (Hu, 2008; Schennach, 2007).

Our work is partly motivated by the elegant work of Hu (2008), who showed that the misclassification probabilities and the latent model for X are nonparametrically identifiable in the presence of a discrete instrumental variable. Hu (2008) then provided a matrix diagonalization technique to estimate the parameters. In contrast to Hu's work on nonparametric identification, we (i) consider the presence of many confounding variables and several instrumental variables, (ii) model the distribution of X parametrically, and (iii) assume that misclassification probabilities do not depend on other variables, and then develop two methods of estimation in the non-trivial conditional likelihood set-up. A parametric model assumption on the conditional distribution of X is necessary to make the method numerically work for a reasonable sample size.

We are proposing two methods of estimation. The first method has two steps. In the first step, the conditional distribution of the true exposure given the confounding variables and instrumental variables and the misclassification probabilities are estimated. In the second step, we obtain an induced model of the response given the observed variables and the estimated parameters from the first step. Next we form a conditional likelihood and maximize this likelihood to estimate the disease–exposure association parameters. The second method is the efficient method. A brief outline of the remainder of the article is as follows. Some background information is given in Section 2, while the details of the methodology are given in Section 3. Section 4 contains simulation studies. Analysis of a real data set that motivated our work is given in Section 5. We formed a nested case–control data from the US birth cohort from the year of 1989, and then analyzed the data for the effect of smoking during pregnancy, the exposure of interest, on the incidence of low birth weight. Since the average number of cigarettes smoked daily cannot be measured accurately, we propose to analyze this data set taking into account instrumental variables. Importantly, for this data set, there are no validation data to access the misclassification probabilities. Finally, we conclude with a discussion in Section 6.

2. Models and background

Suppose that we have an 1:M matched case–control data set with n strata. That means there are one case subject and M control subjects within each stratum. We shall use $a_{i,j}$ to denote the variable a for the jth subject in the ith stratum. Typically, a matched data set is denoted by $Y_{i,j}, X_{i,j}, \boldsymbol{Z}_{i,j}, \boldsymbol{S}_i, j = 1, \ldots, (M + 1), i = 1, \ldots, n$, where Y, X, and $\boldsymbol{Z}$ represent the response, the main binary exposure, and prognostic factors used for adjustment, and $\boldsymbol{S}$ represents a set of confounding variables that are used for matching. Among several definitions we shall stick to the definition that the confounding variables $\boldsymbol{S}$ influence both X and Y (Greenland *and others*, 1999). Since the matching is done based on the values of the confounding variables, all subjects within a matched set would have the same value of $\boldsymbol{S}$. Accordingly, the value of $\boldsymbol{S}$ does not vary within a stratum. To present a more general setting we write prognostic factors $\boldsymbol{Z} = (\boldsymbol{Z}_1^T, \boldsymbol{Z}_2^T)^T$ where $\boldsymbol{Z}_1$ is the prognostic factor that is causally independent of the exposure and $\boldsymbol{Z}_2$ is the prognostic factor that is a confounding variable. In our context, X is never observed, rather W is observed which is a misclassified version of X. We assume that W is independent of $(\boldsymbol{S}, \boldsymbol{X}^*, \boldsymbol{Z}, Y)$ conditional on the true exposure X. Define $\alpha_0 = \text{pr}(W = 1|X = 0) = \text{pr}(W = 1|X = 0, Y = 0) = \text{pr}(W = 1|X = 0, Y = 1)$ and $\alpha_1 = \text{pr}(W = 0|X = 0) = \text{pr}(W = 0|X = 1, Y = 0) = \text{pr}(W = 0|X = 1, Y = 1)$. It is important that we adopt the non-differential misclassification assumption. If the misclassification is differential, then $\text{pr}(W = w|X = x, Y = 0) \neq \text{pr}(W = w|X = x, Y = 1)$ for some (w, x), and that leads to the violation of the conditional independence assumption between W and $(\boldsymbol{S}, \boldsymbol{X}^*, Y, \boldsymbol{Z})$ given X. Additionally, we assume that a set of instrumental variables $\boldsymbol{X}^*$ for X is observed in the data. According to the definition given in Greenland (2000) we assume that (i) $\boldsymbol{X}^*$ directly influences $\boldsymbol{X}$, (ii) $\boldsymbol{X}^*$ may directly influence $\boldsymbol{Z}_1$ (a non-confounder), (iii) $\boldsymbol{X}^*$ is independent of $\boldsymbol{S}$ and $\boldsymbol{Z}_2$ (any confounding variable), and (iv) $\boldsymbol{X}^*$ and Y are

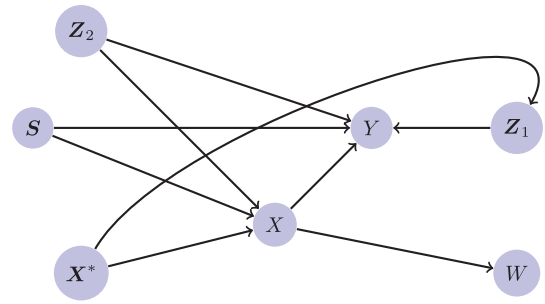

Fig. 1. Schematic diagram to show how variables are related.

independent conditional on X and $\boldsymbol{Z}_1$. It is important to mention that statistical validity of the proposed methods does not rely on the dependence of $\boldsymbol{X}^*$ and any component of $\boldsymbol{Z}$. These assumptions are reflected in Figure 1 where a variable at the end of an arrow is the variable that influences the variable on the head of an arrow.

In our set-up, the observed data are $\{Y_{i,j}, \boldsymbol{X}^*_{i,j}, W_{i,j}, \boldsymbol{Z}_{i,j}, \boldsymbol{S}_i, j = 1, \ldots, (M+1), i = 1, \ldots, n\}$. In general, the assumed model of Y for the ith stratum is

$$\text{pr}(Y_{i,j} = 1 | \boldsymbol{S}_i, X_{i,j}, \boldsymbol{Z}_{i,j}) = H\{g_0(\boldsymbol{S}_i) + \beta_1 X_{i,j} + \boldsymbol{\beta}_2^T \boldsymbol{Z}_{i,j}\}, \tag{2.1}$$

where $H(u) = 1/\{1 + \exp(-u)\}$. Here $g_0(\boldsymbol{S}_i)$ denotes the effect of the stratification variable on the success probability of the response variable, while β_1 and $\boldsymbol{\beta}_2$ are the regression coefficients (association parameters) for X and $\boldsymbol{Z}$, respectively. One can generalize the model by including an interaction term between X and $\boldsymbol{Z}$. In the absence of $\boldsymbol{Z}_2$ that means when $\boldsymbol{S}$ is the only set of confounding variables and they are used for matching, the regression parameter β_1 of Equation (2.1) will have a causal interpretation conditional on $\boldsymbol{Z} = \boldsymbol{Z}_1$ (Hernán and Robins, 2008). Importantly, irrespective of the presence of $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ and the dependence between $\boldsymbol{X}^*$ and $\boldsymbol{Z}_1$ and the independence of $\boldsymbol{X}^*$ and $\boldsymbol{Z}_2$, the proposed estimation techniques are always valid. When X is observed in the data, the parameter $\boldsymbol{\beta} = (\beta_1, \boldsymbol{\beta}_2^T)^T$ is estimated by maximizing the conditional likelihood $\mathcal{L}_c(\boldsymbol{\beta})$ which eliminates the nuisance parameter $g_0(\boldsymbol{S}_i)$ after conditioning on the number of cases for each stratum. Hence,

$$\begin{aligned}
\mathcal{L}_c(\boldsymbol{\beta}|X, \boldsymbol{Z}) &= \prod_{i=1}^{n} \text{pr}\{Y = Y_{i,j}, j = 1, \ldots, (M+1) | \boldsymbol{S}_i, X_{i,j}, \boldsymbol{Z}_{i,j}, j = 1, \ldots, (M+1), \sum_{j=1}^{M+1} Y_{i,j} = 1\} \\
&= \prod_{i=1}^{n} \frac{\prod_{j=1}^{M+1} \text{pr}^{Y_{i,j}}(Y = 1 | \boldsymbol{S}_i, X_{i,j}, \boldsymbol{Z}_{i,j}) \text{pr}^{(1-Y_{i,j})}(Y = 0 | \boldsymbol{S}_i, X_{i,j}, \boldsymbol{Z}_{i,j})}{\sum_{k=1}^{M+1} \text{pr}(Y = 1 | \boldsymbol{S}_i, X_{i,k}, \boldsymbol{Z}_{i,k}) \prod_{r \neq k} \text{pr}(Y = 0 | \boldsymbol{S}_i, X_{i,r}, \boldsymbol{Z}_{i,r})} \\
&= \prod_{i=1}^{n} \frac{\sum_{j=1}^{M+1} Y_{i,j} \exp(\beta_1 X_{i,j} + \boldsymbol{\beta}_2^T \boldsymbol{Z}_{i,j})}{\sum_{j=1}^{M+1} \exp(\beta_1 X_{i,j} + \boldsymbol{\beta}_2^T \boldsymbol{Z}_{i,j})}.
\end{aligned}$$

In the naive approach, X is replaced by W in $\mathcal{L}_c$ and the estimators are defined as $\text{argmax}_{\boldsymbol{\beta}} \mathcal{L}_c(\boldsymbol{\beta}|W, \boldsymbol{Z})$. In principle, the naive estimators are biased, and the degree of bias depends on the severity of the misclassification.

3. Proposed methodology

3.1. *Intuitive estimator*

Our first approach is an intuitive one. Note that the goal is to estimate the regression parameters $\boldsymbol{\beta}$ in Model (2.1), and for that purpose we need the conditional distribution of Y given the observable random variables $\boldsymbol{S}, W, \boldsymbol{X}^*, \boldsymbol{Z}$. To do this, first we specify a model for X given $\boldsymbol{S}, \boldsymbol{X}^*, \boldsymbol{Z}$ among the controls. This model along with the misclassification probabilities induces a model for X given $\boldsymbol{S}, W, \boldsymbol{X}^*, \boldsymbol{Z}$ among the controls denoted by $\text{pr}(X = 1|\boldsymbol{S}, W, \boldsymbol{X}^*, Y = 0, \boldsymbol{Z})$. Second, the resulting induced model $\text{pr}(X = 1|\boldsymbol{S}, W, \boldsymbol{X}^*, Y = 0, \boldsymbol{Z})$ along with Model (2.1) yields a model for Y given $\boldsymbol{S}, W, \boldsymbol{X}^*, \boldsymbol{Z}$. Hence the estimation is carried out in two steps. To proceed, first we assume a parametric model for the probability of $X = 1$ given $\boldsymbol{S}, \boldsymbol{X}^*, \boldsymbol{Z}$ in the control population, and use a logistic model for this probability:

$$\text{pr}(X = 1|\boldsymbol{S}, \boldsymbol{X}^*, Y = 0, \boldsymbol{Z}) = H(\gamma_0 + \boldsymbol{\gamma}_1^T \boldsymbol{S} + \boldsymbol{\gamma}_2^T \boldsymbol{X}^* + \boldsymbol{\gamma}_3^T \boldsymbol{Z}). \tag{3.2}$$

We shall use $H(\boldsymbol{\gamma}, \boldsymbol{S}, \boldsymbol{X}^*, \boldsymbol{Z})$ to denote $H(\gamma_0 + \boldsymbol{\gamma}_1^T \boldsymbol{S} + \boldsymbol{\gamma}_2^T \boldsymbol{X}^* + \boldsymbol{\gamma}_3^T \boldsymbol{Z})$, where $\boldsymbol{\gamma} = (\gamma_0, \boldsymbol{\gamma}_1^T, \boldsymbol{\gamma}_2^T, \boldsymbol{\gamma}_3^T)^T$. Next we obtain the induced conditional probability model for the observed W:

$$\begin{aligned}
&\text{pr}(W = 1|\boldsymbol{S}, \boldsymbol{X}^*, Y = 0, \boldsymbol{Z}) \\
&\quad = \text{pr}(W = 1|\boldsymbol{S}, X = 0, \boldsymbol{X}^*, Y = 0, \boldsymbol{Z})\text{pr}(X = 0|\boldsymbol{S}, \boldsymbol{X}^*, Y = 0, \boldsymbol{Z}) \\
&\qquad + \text{pr}(W = 1|\boldsymbol{S}, X = 1, \boldsymbol{X}^*, Y = 0, \boldsymbol{Z})\text{pr}(X = 1|\boldsymbol{S}, \boldsymbol{X}^*, Y = 0, \boldsymbol{Z}) \\
&\quad = \text{pr}(W = 1|X = 0, Y = 0, \boldsymbol{Z})\text{pr}(X = 0|\boldsymbol{S}, \boldsymbol{X}^*, Y = 0, \boldsymbol{Z}) \\
&\qquad + \text{pr}(W = 1|X = 1, Y = 0, \boldsymbol{Z})\text{pr}(X = 1|\boldsymbol{S}, \boldsymbol{X}^*, Y = 0, \boldsymbol{Z}) \\
&\quad = \alpha_0\{1 - \text{pr}(X = 1|\boldsymbol{S}, \boldsymbol{X}^*, Y = 0, \boldsymbol{Z})\} + (1 - \alpha_1)\text{pr}(X = 1|\boldsymbol{S}, \boldsymbol{X}^*, Y = 0, \boldsymbol{Z}) \\
&\quad = \alpha_0 + (1 - \alpha_0 - \alpha_1)H(\boldsymbol{\gamma}, \boldsymbol{S}, \boldsymbol{X}^*, \boldsymbol{Z}),
\end{aligned} \tag{3.3}$$

where the second equality holds due to our assumptions on the misclassification probability and α_0 and α_1 were defined in the first paragraph of Section 2.

For identifiability of the model parameters, following Hausman *and others* (1998), we assume that $0 < \alpha_0 + \alpha_1 < 1$. A detailed proof of the identifiability is given in Appendix A.1 of the supplementary material available at *Biostatistics* online. Particularly, Hausman *and others* (1998) used this restriction for parameter identification in the case of misclassified dependent variables. We write α_0 and α_1 as $\alpha_0 \equiv \alpha_0(\boldsymbol{\eta}) = \exp(\eta_0)/\{1 + \exp(\eta_0) + \exp(\eta_1)\}$ and $\alpha_1 \equiv \alpha_1(\boldsymbol{\eta}) = \exp(\eta_1)/\{1 + \exp(\eta_0) + \exp(\eta_1)\}$, for $\eta_0, \eta_1 \in \mathcal{R}$ so that the condition $0 < \alpha_0 + \alpha_1 < 1$ is automatically satisfied. Denote $\text{pr}(W = 1|\boldsymbol{S}, \boldsymbol{X}^*, Y = 0, \boldsymbol{Z})$ by $p_w(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{S}, \boldsymbol{X}^*, \boldsymbol{Z})$. Then we propose to estimate $\boldsymbol{\gamma}$ and $\boldsymbol{\eta} = (\eta_0, \eta_1)^T$ by maximizing the following likelihood

$$\mathcal{L}_1(\boldsymbol{\gamma}, \boldsymbol{\eta}) = \prod_{i=1}^{n} \sum_{j=1}^{M+1} \left[\left\{ p_w(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{S}_i, \boldsymbol{X}_{i,j}^*, \boldsymbol{Z}_{i,j}) \right\}^{W_{i,j}} \left\{ 1 - p_w(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{S}_i, \boldsymbol{X}_{i,j}^*, \boldsymbol{Z}_{i,j}) \right\}^{1-W_{i,j}} \right]^{(1-Y_{i,j})}.$$

The estimators $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\eta}}$ for $\boldsymbol{\gamma}$ and $\boldsymbol{\eta}$ are defined as

$$(\boldsymbol{\gamma}, \boldsymbol{\eta}) = \arg\max_{\boldsymbol{\gamma}, \boldsymbol{\eta}} \mathcal{L}_1(\boldsymbol{\gamma}, \boldsymbol{\eta}).$$

More specifically, $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\eta}}$ are obtained by solving $\boldsymbol{S}_{\gamma}(\boldsymbol{\eta},\boldsymbol{\gamma}) = \sum_{i=1}^{n} U_{i,\gamma}(\boldsymbol{\eta},\boldsymbol{\gamma}) = \mathbf{0}$ and $\boldsymbol{S}_{\eta}(\boldsymbol{\eta},\boldsymbol{\gamma}) = \sum_{i=1}^{n} U_{i,\eta}(\boldsymbol{\eta},\boldsymbol{\gamma}) = \mathbf{0}$, where

$$\begin{aligned}
\boldsymbol{U}_{i,\gamma}(\boldsymbol{\gamma},\boldsymbol{\eta}) &= \sum_{j=1}^{M+1}(1-Y_{i,j})\left\{W_{i,j} - p_w(\boldsymbol{\eta},\boldsymbol{\gamma},\boldsymbol{S}_i,\boldsymbol{X}^*_{i,j},\boldsymbol{Z}_{i,j})\right\} \\
&\quad \times \frac{1}{p_w(\boldsymbol{\eta},\boldsymbol{\gamma},\boldsymbol{S}_i,\boldsymbol{X}^*_{i,j},\boldsymbol{Z}_{i,j})\{1-p_w(\boldsymbol{\eta},\boldsymbol{\gamma},\boldsymbol{S}_i,X^*_{i,j},\boldsymbol{Z}_{i,j})\}} \\
&\quad \times \{1-\alpha_0(\boldsymbol{\eta})-\alpha_1(\boldsymbol{\eta})\}H(\boldsymbol{\gamma},\boldsymbol{S}_i,\boldsymbol{X}^*_{i,j},\boldsymbol{Z}_{i,j})\{1-H(\boldsymbol{\gamma},\boldsymbol{S}_i,\boldsymbol{X}^*_{i,j},\boldsymbol{Z}_{i,j})\}\begin{pmatrix}1\\ \boldsymbol{S}_i\\ \boldsymbol{X}^*_{i,j}\\ \boldsymbol{Z}_{i,j}\end{pmatrix}, \\
\boldsymbol{U}_{i,\eta}(\boldsymbol{\gamma},\boldsymbol{\eta}) &= \sum_{j=1}^{M+1}(1-Y_{i,j})\left\{W_{i,j} - p_w(\boldsymbol{\eta},\boldsymbol{\gamma},\boldsymbol{S}_i,\boldsymbol{X}^*_{i,j},\boldsymbol{Z}_{i,j})\right\} \\
&\quad \times \frac{1}{p_w(\boldsymbol{\eta},\boldsymbol{\gamma},\boldsymbol{S}_i,\boldsymbol{X}^*_{i,j},\boldsymbol{Z}_{i,j})\{1-p_w(\boldsymbol{\eta},\boldsymbol{\gamma},\boldsymbol{S}_i,\boldsymbol{X}^*_{i,j},\boldsymbol{Z}_{i,j})\}} \\
&\quad \times \begin{bmatrix} \alpha_0(\boldsymbol{\eta})\{1-\alpha_0(\boldsymbol{\eta})\} - \alpha_0(\boldsymbol{\eta})\{1-\alpha_0(\boldsymbol{\eta})-\alpha_1(\boldsymbol{\eta})\}H(\boldsymbol{\gamma},\boldsymbol{S}_i,\boldsymbol{X}^*_{i,j},\boldsymbol{Z}_{i,j}) \\ -\alpha_0(\boldsymbol{\eta})\alpha_1(\boldsymbol{\eta}) - \alpha_1(\boldsymbol{\eta})\{1-\alpha_0(\boldsymbol{\eta})-\alpha_1(\boldsymbol{\eta})\}H(\boldsymbol{\gamma},\boldsymbol{S}_i,\boldsymbol{X}^*_{i,j},\boldsymbol{Z}_{i,j}) \end{bmatrix}.
\end{aligned}$$

We then obtain the induced model for X given $W=\omega, \boldsymbol{X}^*, \boldsymbol{S}$, and $\boldsymbol{Z}$ among the control subjects as follows:

$$\begin{aligned}
\mathrm{pr}(X=1|\boldsymbol{S},W=1,\boldsymbol{X}^*,Y=0,\boldsymbol{Z}) &= \frac{\mathrm{pr}(W=1|\boldsymbol{S},X=1,\boldsymbol{X}^*,Y=0,\boldsymbol{Z})\mathrm{pr}(X=1|\boldsymbol{S},\boldsymbol{X}^*,Y=0,\boldsymbol{Z})}{\mathrm{pr}(W=1|\boldsymbol{S},\boldsymbol{X}^*,Y=0,\boldsymbol{Z})} \\
&= \frac{(1-\alpha_1)H(\boldsymbol{\gamma},\boldsymbol{S},\boldsymbol{X}^*,\boldsymbol{Z})}{\alpha_0+(1-\alpha_0-\alpha_1)H(\boldsymbol{\gamma},\boldsymbol{S},\boldsymbol{X}^*,\boldsymbol{Z})}, \qquad (3.4) \\
\mathrm{pr}(X=1|\boldsymbol{S},W=0,\boldsymbol{X}^*,Y=0,\boldsymbol{Z}) &= \frac{\mathrm{pr}(W=0|\boldsymbol{S},X=1,\boldsymbol{X}^*,Y=0,\boldsymbol{Z})\mathrm{pr}(X=1|\boldsymbol{S},\boldsymbol{X}^*,Y=0,\boldsymbol{Z})}{\mathrm{pr}(W=0|\boldsymbol{S},\boldsymbol{X}^*,Y=0,\boldsymbol{Z})} \\
&= \frac{\alpha_1 H(\boldsymbol{\gamma},\boldsymbol{S},\boldsymbol{X}^*,\boldsymbol{Z})}{1-\alpha_0-(1-\alpha_0-\alpha_1)H(\boldsymbol{\gamma},\boldsymbol{S},\boldsymbol{X}^*,\boldsymbol{Z})}. \qquad (3.5)
\end{aligned}$$

Next, we obtain the induced model for the response Y given $\boldsymbol{S}$, W, $\boldsymbol{X}^*$, $\boldsymbol{Z}$ that is stated in the following lemma whose proof is given in Appendix A.2 of the supplementary material available at *Biostatistics* online.

LEMMA 1 Under the assumptions stated previously the induced model for Y given $\boldsymbol{S}$, W, $\boldsymbol{X}^*$, and $\boldsymbol{Z}$ is

$$\mathrm{pr}(Y=1|\boldsymbol{S},W,\boldsymbol{X}^*,\boldsymbol{Z}) = H\{g_0(\boldsymbol{S}) + \boldsymbol{\beta}_2^T\boldsymbol{Z} + g_1(\beta_1,\boldsymbol{S}_i,W,\boldsymbol{X}^*,\boldsymbol{Z},\boldsymbol{\eta},\boldsymbol{\gamma})\}, \qquad (3.6)$$

where

$$\exp\{g_1(\beta_1,\boldsymbol{S},W=1,\boldsymbol{X}^*,\boldsymbol{Z},\boldsymbol{\gamma},\boldsymbol{\eta})\} = \frac{\exp(\beta_1)(1-\alpha_1)H(\boldsymbol{\gamma},\boldsymbol{S},\boldsymbol{X}^*,\boldsymbol{Z}) + \alpha_0\{1-H(\boldsymbol{\gamma},\boldsymbol{S},\boldsymbol{X}^*,\boldsymbol{Z})\}}{\alpha_0+(1-\alpha_0-\alpha_1)H(\boldsymbol{\gamma},\boldsymbol{S},\boldsymbol{X}^*,\boldsymbol{Z})},$$

$$\exp\{g_1(\beta_1,\boldsymbol{S},W=0,\boldsymbol{X}^*,\boldsymbol{Z},\boldsymbol{\gamma},\boldsymbol{\eta})\} = \frac{\exp(\beta_1)\alpha_1 H(\boldsymbol{\gamma},\boldsymbol{S},\boldsymbol{X}^*,\boldsymbol{Z}) + (1-\alpha_0)\{1-H(\boldsymbol{\gamma},\boldsymbol{S},\boldsymbol{X}^*,\boldsymbol{Z})\}}{1-\alpha_0-(1-\alpha_0-\alpha_1)H(\boldsymbol{\gamma},\boldsymbol{S},\boldsymbol{X}^*,\boldsymbol{Z})}.$$

Model (3.6) for the response Y is in terms of observed variables, $\boldsymbol{S}$, W, $\boldsymbol{X}^*$, and $\boldsymbol{Z}$, and it involves the main association parameters $\boldsymbol{\beta}$. To estimate $\boldsymbol{\beta}$, we form the conditional likelihood function based on this induced probability model and maximize with respect to $\boldsymbol{\beta}$ to estimate the parameter. Define the estimator

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \mathcal{L}_2(\boldsymbol{\beta}|\hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\gamma}}),$$

where the conditional likelihood function is

$$\begin{aligned}
\mathcal{L}_2(\boldsymbol{\beta}|\boldsymbol{\eta},\boldsymbol{\gamma}) &= \prod_{i=1}^{n} \mathrm{pr}\{Y_{i,j}, j=1,\ldots,(M+1)|\boldsymbol{S}_i, W_{i,j}, \boldsymbol{X}^*_{i,j}, \boldsymbol{Z}_{i,j}, j=1,\ldots,(M+1), \sum_{j=1}^{M+1} Y_{i,j}=1\} \\
&= \prod_{i=1}^{n} \frac{\prod_{j=1}^{M+1} \mathrm{pr}^{Y_{i,j}}(Y=1|\boldsymbol{S}_i, W_{i,j}, \boldsymbol{X}^*_{i,j}, \boldsymbol{Z}_{i,j}) \mathrm{pr}^{(1-Y_{i,j})}(Y=0|\boldsymbol{S}_i, W_{i,j}, \boldsymbol{X}^*_{i,j}, \boldsymbol{Z}_{i,j})}{\sum_{k=1}^{M+1} \mathrm{pr}(Y=1|\boldsymbol{S}_i, W_{i,k}, \boldsymbol{X}^*_{i,k}, \boldsymbol{Z}_{i,k}) \prod_{r\neq k} \mathrm{pr}(Y=0|\boldsymbol{S}_i, W_{i,r}, \boldsymbol{X}^*_{i,r}, \boldsymbol{Z}_{i,r})} \\
&= \prod_{i=1}^{n} \frac{\sum_{j=1}^{M+1} Y_{i,j} \exp\{\boldsymbol{\beta}_2^T \boldsymbol{Z}_{i,j} + g_1(\beta_1, \boldsymbol{S}_i, W_{i,j}, \boldsymbol{X}^*_{i,j}, \boldsymbol{Z}_{i,j}, \boldsymbol{\eta}, \boldsymbol{\gamma})\}}{\sum_{j=1}^{M+1} \exp\{\boldsymbol{\beta}_2^T \boldsymbol{Z}_{i,j} + g_1(\beta_1, \boldsymbol{S}_i, W_{i,j}, \boldsymbol{X}^*_{i,j}, \boldsymbol{Z}_{i,j}, \boldsymbol{\gamma}, \boldsymbol{\eta})\}} \\
&= \prod_{i=1}^{n} \frac{\exp[\sum_{j=1}^{M+1} Y_{i,j}\{\boldsymbol{\beta}_2^T \boldsymbol{Z}_{i,j} + g_1(\beta_1, \boldsymbol{S}_i, W_{i,j}, \boldsymbol{X}^*_{i,j}, \boldsymbol{Z}_{i,j}, \boldsymbol{\eta}, \boldsymbol{\gamma})]}{\sum_{j=1}^{M+1} \exp\{\boldsymbol{\beta}_2^T \boldsymbol{Z}_{i,j} + g_1(\beta_1, \boldsymbol{S}_i, W_{i,j}, \boldsymbol{X}^*_{i,j}, \boldsymbol{Z}_{i,j}, \boldsymbol{\eta}, \boldsymbol{\gamma})\}}.
\end{aligned}$$

Therefore, $\hat{\boldsymbol{\beta}}$ is obtained by solving $S_{\beta_1}(\boldsymbol{\beta}, \hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\gamma}}) = \sum_{i=1}^{n} U_{i,\beta_1}(\boldsymbol{\beta}, \hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\gamma}}) = 0$, $\boldsymbol{S}_{\boldsymbol{\beta}_2}(\boldsymbol{\beta}, \hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\gamma}}) = \sum_{i=1}^{n} \boldsymbol{U}_{i,\boldsymbol{\beta}_2}(\boldsymbol{\beta}, \hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\gamma}}) = \boldsymbol{0}$, where

$$\begin{aligned}
U_{i,\beta_1}(\boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\gamma}) &= \sum_{j=1}^{M+1} \left[Y_{i,j} - \frac{\exp\{\boldsymbol{\beta}_2^T \boldsymbol{Z}_{i,j} + g_1(\beta_1, \boldsymbol{S}_i, W_{i,j}, \boldsymbol{X}^*_{i,j}, \boldsymbol{Z}_{i,j}, \boldsymbol{\eta}, \boldsymbol{\gamma})\}}{\sum_{k=1}^{M+1} \exp\{\boldsymbol{\beta}_2^T \boldsymbol{Z}_{i,k} + g_1(\beta_1, \boldsymbol{S}_i, W_{i,k}, \boldsymbol{X}^*_{i,k}, \boldsymbol{Z}_{i,j}, \boldsymbol{\eta}, \boldsymbol{\gamma})\}} \right], \\
&\quad \times g_{\beta_1}(\beta_1, \boldsymbol{S}_i, W_{i,j}, \boldsymbol{X}^*_{i,j}, \boldsymbol{Z}_{i,j}, \boldsymbol{\eta}, \boldsymbol{\gamma}), \\
\boldsymbol{U}_{i,\boldsymbol{\beta}_2}(\boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\gamma}) &= \sum_{j=1}^{M+1} \left[Y_{i,j} - \frac{\exp\{\boldsymbol{\beta}_2^T \boldsymbol{Z}_{i,j} + g_1(\beta_1, \boldsymbol{S}_i, W_{i,j}, \boldsymbol{X}^*_{i,j}, \boldsymbol{Z}_{i,j}, \boldsymbol{\eta}, \boldsymbol{\gamma})\}}{\sum_{k=1}^{M+1} \exp\{\boldsymbol{\beta}_2^T \boldsymbol{Z}_{i,k} + g_1(\beta_1, \boldsymbol{S}_i, W_{i,k}, \boldsymbol{X}^*_{i,k}, \boldsymbol{Z}_{i,j}, \boldsymbol{\eta}, \boldsymbol{\gamma})\}} \right] \boldsymbol{Z}_{i,j},
\end{aligned}$$

with $g_{\beta_1}(\cdot) = \partial g_1(\cdot)/\partial \beta_1$. We now present the following main theorem. Its proof is given in Appendix A.3 of the supplementary material available at *Biostatistics* online.

THEOREM 1 Under standard regularity conditions and as $n \to \infty$, the asymptotic distribution of $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ converges to a mean-zero normal distribution, and the asymptotic variance of $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ can be consistently estimated by the last $(p+1)$ rows and the last $(p+1)$ columns of $\hat{A}^{-1}(\sum_{i=1}^{n} \boldsymbol{U}_i \boldsymbol{U}_i^T/n)\hat{A}^{-T}$, where $\hat{A} = -(1/n)\partial \boldsymbol{S}_\theta/\partial \boldsymbol{\theta}$, $\boldsymbol{S}_\theta = (\boldsymbol{S}_\gamma^T(\boldsymbol{\gamma}, \boldsymbol{\eta}), \boldsymbol{S}_\eta^T(\boldsymbol{\gamma}, \boldsymbol{\eta}), S_{\beta_1}(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\eta}), \boldsymbol{S}_{\boldsymbol{\beta}_2}^T(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\eta}))^T$ and $\boldsymbol{U}_i = (\boldsymbol{U}_{i,\gamma}^T, \boldsymbol{U}_{i,\eta}^T, U_{i,\beta_1}, \boldsymbol{U}_{i,\boldsymbol{\beta}_2}^T)^T$.

3.2. *Efficient estimator*

For deriving the efficient estimator, we need the following lemma which is proved in Appendix A.4 of the supplementary material available at *Biostatistics* online.

LEMMA 2 Under the assumed Models (2.1) and (3.2) and the assumptions on the misclassification probabilities, we obtain

$$(i)\ \ \mathrm{pr}(Y=1|\boldsymbol{S},\boldsymbol{X}^*,\boldsymbol{Z}) = H\{g_0(\boldsymbol{S}) + \boldsymbol{\beta}_2^T\boldsymbol{Z} + g_2(\boldsymbol{\gamma},\beta_1,\boldsymbol{S},\boldsymbol{X}^*,\boldsymbol{Z})\},$$
$$(ii)\ \ \mathrm{pr}(X=1|\boldsymbol{S},\boldsymbol{X}^*,\boldsymbol{Z},Y=1) = H(\gamma_0+\beta_1+\boldsymbol{\gamma}_1^T\boldsymbol{S}+\boldsymbol{\gamma}_2^T\boldsymbol{X}^*+\boldsymbol{\gamma}_3^T\boldsymbol{Z}),$$
$$(iii)\ \ \mathrm{pr}(W=1|\boldsymbol{S},\boldsymbol{X}^*,\boldsymbol{Z},Y=1) = \alpha_0+(1-\alpha_0-\alpha_1)H(\gamma_0+\beta_1+\boldsymbol{\gamma}_1^T\boldsymbol{S}+\boldsymbol{\gamma}_2^T\boldsymbol{X}^*+\boldsymbol{\gamma}_3^T\boldsymbol{Z}),$$

where

$$g_2(\boldsymbol{\gamma},\beta_1,\boldsymbol{S},\boldsymbol{X}^*,\boldsymbol{Z}) = \log\{1-H(\boldsymbol{\gamma},\boldsymbol{S},\boldsymbol{X}^*,\boldsymbol{Z})+\exp(\beta_1)H(\boldsymbol{\gamma},\boldsymbol{S},\boldsymbol{X}^*,\boldsymbol{Z})\}.$$

The likelihood of the observed data $\{W_{i,j}, Y_{i,j}, i=1,\ldots,(M+1)\}$ from the ith stratum conditional on $\boldsymbol{S}_i,\boldsymbol{X}^*_{i,j},\boldsymbol{Z}_{i,j},j=1,\ldots,(M+1)$ is

$$\mathcal{L}_i = \prod_{j=1}^{M+1}\mathrm{pr}(W_{i,j},Y_{i,j}|\boldsymbol{S}_i,\boldsymbol{X}^*_{i,j},\boldsymbol{Z}_{i,j}) = \prod_{j=1}^{M+1}\mathrm{pr}(W_{i,j}|\boldsymbol{S}_i,\boldsymbol{X}^*_{i,j},Y_{i,j},\boldsymbol{Z}_{i,j})\mathrm{pr}(Y_{i,j}|\boldsymbol{S}_i,\boldsymbol{X}^*_{i,j},\boldsymbol{Z}_{i,j}).$$

Lemma 2 indicates that $\mathrm{pr}(Y_{i,j}|\boldsymbol{S}_i,\boldsymbol{X}^*_{i,j},\boldsymbol{Z}_{i,j})$ is in the logistic form that still involves with the nuisance intercept parameter $g_0(\boldsymbol{S}_i)$, and its dimension increases with the number of matched sets n. However, the complete sufficient statistic for the stratum specific intercept $g_0(\boldsymbol{S}_i)$ is $\sum_{j=1}^{M+1}Y_{i,j}$, the total number of successes in the ith stratum. Conditioning on the number of successes results in a conditional likelihood that is free from this stratum specific intercept. Following the arguments of Godambe (1976) and Rathouz *and others* (2002), the maximum conditional likelihood estimator of $\boldsymbol{\theta}=(\boldsymbol{\beta}^T,\boldsymbol{\gamma}^T,\eta_0,\eta_1)^T$ (η_0,η_1 in lieu of α_0 and α_1) is then semiparametric efficient. The conditional likelihood for the ith stratum is

$$\begin{aligned}
\mathcal{L}_{i,c} &= \left\{\prod_{j=1}^{M+1}\mathrm{pr}(W_{i,j}|\boldsymbol{S}_i,\boldsymbol{X}^*_{i,j},Y_{i,j},\boldsymbol{Z}_{i,j})\right\}\\
&\quad\times\mathrm{pr}(Y_{i,1},\ldots,Y_{i,M+1}|\boldsymbol{S}_i,\boldsymbol{X}^*_{i,j},\boldsymbol{Z}_{i,j},j=1,\ldots,(M+1),\sum_{j=1}^{M+1}Y_{i,j}=1)\\
&= \prod_{j=1}^{M+1}\left\{\mathrm{pr}^{W_{i,j}}(W_{i,j}=1|\boldsymbol{S}_i,\boldsymbol{X}^*_{i,j},Y_{i,j}=1,\boldsymbol{Z}_{i,j})\mathrm{pr}^{(1-W_{i,j})}(W_{i,j}=0|\boldsymbol{S}_i,\boldsymbol{X}^*_{i,j},Y_{i,j}=1,\boldsymbol{Z}_{i,j})\right\}^{Y_{i,j}}\\
&\quad\times\left\{\mathrm{pr}^{W_{i,j}}(W_{i,j}=1|\boldsymbol{S}_i,\boldsymbol{X}^*_{i,j},Y_{i,j}=0,\boldsymbol{Z}_{i,j})\mathrm{pr}^{(1-W_{i,j})}(W_{i,j}=0|\boldsymbol{S}_i,\boldsymbol{X}^*_{i,j},Y_{i,j}=0,\boldsymbol{Z}_{i,j})\right\}^{1-Y_{i,j}}\\
&\quad\times\frac{\sum_{j=1}^{M+1}Y_{i,j}\exp\{\boldsymbol{\beta}_2^T\boldsymbol{Z}_{i,j}+g_2(\boldsymbol{\gamma},\beta_1,\boldsymbol{S}_i,\boldsymbol{X}^*_{i,j},\boldsymbol{Z}_{i,j})\}}{\sum_{j=1}^{M+1}\exp\{\boldsymbol{\beta}_2^T\boldsymbol{Z}_{i,j}+g_2(\boldsymbol{\gamma},\beta_1,\boldsymbol{S}_i,\boldsymbol{X}^*_{i,j},\boldsymbol{Z}_{i,j})\}}.
\end{aligned}$$

Following Equation (3.3) and Lemma 2, we can replace $\mathrm{pr}(W=1|\boldsymbol{S},\boldsymbol{X}^*,Y=0,\boldsymbol{Z})$ by $\alpha_0+(1-\alpha_0-\alpha_1)H(\gamma_0+\boldsymbol{\gamma}_1^T\boldsymbol{S}+\boldsymbol{\gamma}_2^T\boldsymbol{X}^*+\boldsymbol{\gamma}_3^T\boldsymbol{Z})$ and $\mathrm{pr}(W=1|\boldsymbol{S},\boldsymbol{X}^*,Y=1,\boldsymbol{Z})$ by $\alpha_0+(1-\alpha_0-\alpha_1)H(\gamma_0+\beta_1+\boldsymbol{\gamma}_1^T\boldsymbol{S}+\boldsymbol{\gamma}_2^T\boldsymbol{X}^*+\boldsymbol{\gamma}_3^T\boldsymbol{Z})$. Hence, the efficient estimator $\hat{\boldsymbol{\theta}}_{\mathrm{eff}}$ for $\boldsymbol{\theta}$ is obtained by solving $S_{\mathrm{eff},\theta}=\sum_{i=1}^{n}\partial\log(\mathcal{L}_{i,c})/\partial\boldsymbol{\theta}=\boldsymbol{0}$. The asymptotic standard error of $\hat{\boldsymbol{\theta}}_{\mathrm{eff}}$ can be estimated by inverting the negative of the Hessian matrix $-\sum_{i=1}^{n}\partial^2\log(\mathcal{L}_{i,c})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T$.

4. Simulation study

4.1. *Simulation design*

To simulate a matched case–control data set, we first simulated a large population with size $N = 80\,000$ by generating six variables, (S, W, X, X^*, Z, Y). From that population we sampled n cases ($Y = 1$), and corresponding to each sampled case we sampled $M = 2$ controls randomly from the population after matching the value of the stratification variable S.

The stratification variable S, instrument X^*, and prognostic factor Z were generated from Uniform$(-1, 1)$, Normal$(0, 0.5^2)$, and Normal$(0, 0.5^2)$ distributions, respectively. The exposure X was simulated from the Bernoulli distribution with the success probability $H(\gamma_0 + \gamma_1 S + \gamma_2 X^*)$, $\gamma_0 = -1$, $\gamma_1 = 1$, and we took two different values for γ_2, 1 and 2, corresponding to, respectively, a moderate and a strong association between the exposure and its instrument. This resulted in approximately 30% marginal prevalence of X. In the end, the binary response variable Y was simulated from the Bernoulli distribution with the success probability $H(-2 - 2S + X + 0.5Z)$, so $\beta_1 = 1$ and $\beta_2 = 0.5$, and this resulted in 20% marginal prevalence of $Y = 1$ in the population. We set the surrogate (misclassified) variable $W = B \times X + (1 - B^*) \times (1 - X)$, $B \sim$ Bernoulli$(1 - \alpha_1)$ and $B^* \sim$Bernoulli$(1 - \alpha_0)$. We considered two cases, one where $\alpha_0 = \alpha_1$ and the other where $\alpha_0 \neq \alpha_1$. In the first case, we considered three different misclassification probabilities MC1: $\alpha_0 = \alpha_1 = 0.2$, MC2: $\alpha_0 = \alpha_1 = 0.1$, and MC3: $\alpha_0 = \alpha_1 = 0.05$. In the second case, we considered MC1: $\alpha_0 = 0.2, \alpha_1 = 0.1$; MC2: $\alpha_0 = 0.2, \alpha_1 = 0.05$; MC3: $\alpha_0 = 0.1, \alpha_1 = 0.05$.

For a given case, a control was considered to be matched if the absolute difference between the values of the confounding variable for the case and control subjects was less than 0.01. For a given case, two controls were randomly chosen from the set of all matched controls identified in the population. We considered two different sample sizes, $n = 200$ and 1000. Thus when $n = 200$, this means there are 200 cases and 400 controls in each matched data set, while for $n = 1000$ there are 1000 cases and 2000 controls in each matched data set.

4.2. *Method of analysis*

Each simulated data set was analyzed by four approaches. First, we estimated $\boldsymbol{\beta} = (\beta_1, \beta_2)^T$ using the true X, and call it M1. This method is presented to compare the performance of the other approaches. In a real data analysis, since X is not observed, M1 cannot be used. In the second method, we replaced X by W in the conditional likelihood $\mathcal{L}_c(\boldsymbol{\beta}|X, \boldsymbol{Z})$, refer to this naive method as M2. Third, we analyzed the simulated data sets using the proposed intuitive method (two step estimation), and we refer to it as M3. Finally, we analyzed the data sets by the efficient method, and call it by M4. We conducted 5000 simulations for each one of the four scenarios (two values for n, and two different associations between the instrument and the exposure variable) under different misclassification probabilities. It is important to point out that in the absence of any validation data set or replications, one cannot use the regression calibration approach that is commonly used for reducing bias.

The methods were compared in terms of performance of the estimators of $\boldsymbol{\beta} = (\beta_1, \beta_2)^T$. We present the relative median bias, a robust standard deviation calculated as $(Q3 - Q1)/1.349$ (denoted as SD* in Tables 1–3), where Q_1 and Q_3 are the first and third quartile of the 1000 estimates of the parameter, standard deviation of the estimator based on the 1000 estimates, an average of the estimated standard errors, the 95% coverage probability using the Wald confidence interval, and the mean squared error (MSE). These summary measures were calculated based on the converged data sets. In the computation when either the absolute value of $\hat{\beta}_1$ or the standard error of β_1 is greater than 5, we declare that data set as non-convergent. When $n = 200$, approximately 7–8% data sets did not converge for M3, while 1–2% data sets faced convergence issue in M4. When $n = 1000$, approximately 4–5% data sets did not converge

Table 1. Results of the simulation study under equal misclassification

		MC1								MC2						MC3					
MT		M1		M2		M3		M4		M2		M3		M4		M2		M3		M4	
		β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
$n = 200$																					
S1	B	0.42	−0.25	−53.73	−2.42	−59.75	−0.47	7.12	2.11	−31.67	−1.28	−34.97	0.66	10.17	1.26	−17.72	−1.27	−20.35	1.48	12.11	1.39
	*SD**	22.26	18.23	18.89	18.24	47.53	19.51	64.97	21.90	20.01	18.18	57.68	19.70	47.50	19.74	20.26	18.27	62.18	19.69	39.17	19.21
	SD	22.25	18.72	18.70	18.43	56.18	20.18	82.86	25.48	19.92	18.53	57.82	20.52	59.68	21.31	21.02	18.62	59.35	20.50	48.64	20.04
	SE	21.90	18.48	18.66	18.06	78.78	24.55	63.68	23.14	19.79	18.22	71.11	25.27	49.20	21.60	20.66	18.33	67.07	23.93	42.38	22.41
	CP	94.95	94.61	18.30	94.41	67.19	95.95	85.45	93.20	63.81	94.50	77.65	95.67	91.67	93.71	85.72	94.82	81.26	95.68	93.74	93.72
	MSE	4.96	3.51	32.43	3.40	61.49	4.10	73.52	7.44	13.86	3.43	44.61	4.25	39.55	4.88	7.39	3.47	40.81	4.18	27.32	4.20
S2	B	0.53	0.99	−51.07	−1.93	−41.21	−1.17	1.34	0.60	−29.44	−0.97	−18.21	0.56	3.49	0.95	−15.92	−0.04	−7.54	1.12	4.11	1.21
	*SD**	21.09	18.32	18.48	17.89	39.79	19.36	40.17	19.48	19.15	18.05	37.62	18.79	30.77	18.64	20.33	17.88	32.92	19.15	27.33	18.55
	SD	21.43	18.75	18.56	18.19	42.51	19.94	46.85	21.07	19.29	18.41	38.34	19.51	32.55	19.22	20.27	18.55	36.23	19.38	28.30	18.94
	SE	21.15	18.53	18.50	18.08	56.23	21.88	39.81	21.23	19.44	18.25	45.23	21.23	32.25	21.76	20.16	18.37	38.96	20.23	28.90	20.44
	CP	95.22	94.68	20.47	94.64	72.54	95.01	90.62	93.83	66.64	94.60	83.99	95.17	93.63	94.25	86.83	94.80	89.13	95.02	94.39	94.12
	MSE	4.60	3.52	29.72	3.31	31.54	3.97	22.28	4.46	12.28	3.39	18.26	3.81	10.84	3.71	6.54	3.44	14.07	3.75	8.33	3.60
$n = 1000$																					
S1	B	0.17	0.04	−53.87	−2.59	−22.19	−0.84	1.21	-0.15	−31.75	−1.62	−2.51	0.52	1.69	0.10	−17.64	−0.94	4.27	1.15	2.31	0.19
	*SD**	9.69	8.40	8.41	8.20	38.90	9.13	26.14	8.97	8.62	8.27	27.63	9.00	20.49	8.54	9.09	8.35	24.30	8.85	18.16	8.39
	SD	9.83	8.32	8.33	8.17	36.14	9.42	28.20	8.81	8.83	8.22	28.59	9.05	21.14	8.44	9.25	8.25	27.07	8.93	18.09	8.36
	SE	9.71	8.21	8.30	8.02	41.00	10.57	24.89	9.66	8.79	8.09	32.33	9.76	20.52	9.55	9.17	8.14	26.73	10.09	18.79	9.40
	CP	94.83	94.89	0.04	94.37	73.76	95.50	89.82	94.57	4.98	94.68	81.63	95.56	87.85	94.24	51.33	94.88	85.05	95.49	89.30	94.04
	MSE	0.97	0.69	29.70	0.68	17.66	0.89	7.96	0.78	10.83	0.68	8.22	0.82	4.50	0.72	3.92	0.68	7.56	0.80	3.39	0.71
S2	B	0.09	0.02	−51.24	−2.66	−13.11	−0.88	0.19	−0.30	−29.54	−1.72	−2.74	0.00	0.28	−0.20	−16.03	−0.91	−0.18	0.18	0.55	−0.24
	*SD**	9.19	8.28	8.18	8.12	22.34	8.85	16.99	8.58	8.80	8.15	15.61	8.53	13.92	8.29	8.85	8.23	13.77	8.42	12.36	8.26
	SD	9.54	8.29	8.34	8.09	22.47	8.91	17.74	8.60	8.79	8.15	16.55	8.51	14.13	8.34	9.07	8.21	16.09	8.42	12.34	8.25
	SE	9.38	8.22	8.23	8.03	24.14	9.46	16.10	8.45	8.64	8.10	17.48	9.07	14.00	8.57	8.95	8.15	14.98	8.47	13.27	9.44
	CP	94.36	95.08	0.00	94.44	81.92	95.14	93.10	94.61	7.80	94.76	90.35	95.24	93.38	94.65	56.33	94.88	91.34	94.88	94.34	94.34
	MSE	0.91	0.69	27.04	0.67	7.19	0.79	3.15	0.74	9.47	0.67	2.83	0.73	2.00	0.70	3.39	0.68	2.59	0.71	1.53	0.69

MT, method; B, relative median bias $\times 100$; SD*, simulation standard deviation based on quantiles $\times 100$; SD, simulation standard deviation $\times 100$; SE, estimated standard error $\times 100$; CP, 95% coverage probability based on the Wald confidence interval; MSE, mean squared error; M1, conditional logistic analysis when true X is used; M2, conditional logistic analysis when X is replaced by W; M3, proposed two-step method; M4, proposed efficient estimator; MC1, $\alpha_0 = \alpha_1 = 0.2$; MC2, $\alpha_0 = \alpha_1 = 0.1$; MC3, $\alpha_0 = \alpha_1 = 0.05$, $\text{pr}(X = 1|S, X^*) = H(-1 + S + \gamma_2 X^*)$; S1, $\gamma_2 = 1$; S2, $\gamma_2 = 2$, $\text{pr}(Y = 1|S, X, Z) = H(-2 - 2S + X + 0.5Z)$, $\alpha_0 = \text{pr}(W = 1|X = 0)$, $\alpha_1 = \text{pr}(W = 0|X = 1)$.

Table 2. Results of the simulation study under unequal misclassification

MT		MC1 M1		MC1 M2		MC1 M3		MC1 M4		MC2 M2		MC2 M3		MC2 M4		MC3 M2		MC3 M3		MC3 M4	
		β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
$n = 200$																					
S1	B	0.39	−0.26	−47.20	−1.93	−48.22	0.21	8.87	1.67	−43.91	−1.77	−42.71	0.28	10.33	1.83	−28.53	−1.37	−28.40	0.82	11.10	1.65
	*SD**	22.28	18.24	18.68	18.31	51.93	19.88	57.30	20.61	18.54	18.36	52.82	19.66	54.69	20.29	19.55	18.34	58.46	19.65	44.95	19.50
	SD	22.25	18.72	18.61	18.47	56.22	20.27	71.53	23.31	18.49	18.49	56.68	20.06	68.58	22.59	19.79	18.55	58.85	20.46	55.63	20.64
	SE	21.90	18.48	18.54	18.12	75.03	23.97	57.93	23.41	18.51	18.14	58.27	22.19	54.87	22.94	19.66	18.25	60.48	23.25	47.34	22.72
	CP	94.95	94.61	28.55	94.49	70.98	95.59	87.61	93.72	34.30	94.41	74.19	95.89	88.33	93.51	68.95	94.61	78.57	95.20	91.96	94.10
	MSE	4.96	3.51	25.60	3.41	50.83	4.09	55.76	5.97	22.51	3.42	47.61	4.08	51.93	5.28	11.99	3.44	42.69	4.19	34.68	4.55
S2	B	0.53	0.89	−44.02	−1.43	−29.97	−0.49	3.06	0.97	−40.83	−1.09	−23.31	0.36	3.55	0.81	−25.92	−0.57	−12.76	0.91	3.81	1.02
	*SD**	21.08	18.31	18.60	17.90	39.34	19.45	36.23	19.27	18.61	17.91	39.64	19.45	34.34	19.01	19.16	17.87	35.94	18.91	29.23	18.50
	SD	21.43	18.76	18.44	18.26	39.98	19.86	39.44	19.95	18.49	18.30	40.45	19.77	36.70	19.66	19.21	18.44	37.49	19.47	30.88	19.14
	SE	21.15	18.53	18.40	18.14	52.46	22.56	35.98	21.75	18.39	18.17	42.85	21.80	34.25	21.43	19.34	18.29	42.48	22.05	31.32	21.04
	CP	95.22	94.68	32.97	94.70	78.31	94.98	91.63	93.92	40.42	94.74	80.39	95.08	92.01	94.18	71.98	94.63	86.48	95.22	93.53	94.19
	MSE	4.60	3.52	22.92	3.34	23.57	3.91	15.83	3.99	19.90	3.35	21.35	3.89	13.78	3.87	10.45	3.40	16.14	3.79	9.85	3.66
$n = 1000$																					
S1	B	0.17	0.04	−47.34	−2.15	−11.89	−0.53	0.96	−0.15	−44.07	−2.04	−6.44	−0.17	1.96	−0.10	−28.78	−1.48	1.66	0.71	2.40	0.28
	*SD**	9.69	8.40	8.35	8.19	35.52	9.15	23.67	8.79	8.32	8.19	33.76	9.20	22.94	8.71	8.62	8.31	26.86	8.94	19.84	8.49
	SD	9.83	8.32	8.29	8.19	33.47	9.24	25.54	8.61	8.27	8.20	31.60	9.17	24.51	8.56	8.78	8.24	26.87	9.02	20.32	8.40
	SE	9.71	8.21	8.24	8.04	36.78	9.88	23.03	11.61	8.23	8.06	34.39	9.95	21.52	10.34	8.73	8.10	30.32	9.55	20.06	8.86
	CP	94.84	94.88	0.06	94.54	78.15	95.65	88.47	94.19	0.08	94.64	80.30	95.41	87.16	94.09	9.54	94.74	82.84	95.51	87.46	94.37
	MSE	0.97	0.69	23.07	0.68	12.73	0.86	6.53	0.74	20.02	0.68	10.55	0.84	6.03	0.73	8.99	0.68	7.25	0.81	4.19	0.72
S2	B	0.10	0.01	−44.44	−2.39	−7.25	−0.47	−0.04	−0.10	−40.83	−2.15	−4.05	−0.17	0.38	−0.13	−26.45	−1.47	−0.86	0.00	0.58	−0.24
	*SD**	9.19	8.28	8.15	8.14	18.86	8.73	15.65	8.4	8.15	8.14	17.53	8.66	14.77	8.35	8.58	8.17	14.96	8.43	13.05	8.35
	SD	9.54	8.29	8.27	8.11	19.71	8.74	16.56	8.44	8.23	8.13	18.74	8.68	16.03	8.39	8.72	8.17	15.53	8.47	13.62	8.30
	SE	9.38	8.22	8.18	8.05	20.41	19.77	15.27	9.25	8.18	8.07	18.52	10.33	15.52	11.18	8.59	8.12	16.46	8.97	13.70	9.08
	CP	94.36	95.06	0.06	94.48	86.47	95.37	92.53	94.38	0.16	94.56	87.46	95.27	91.98	94.19	14.28	94.80	92.08	95.33	92.91	94.26
	MSE	0.91	0.69	20.35	0.67	4.53	0.76	2.74	0.71	17.32	0.67	3.76	0.75	2.57	0.70	7.68	0.67	2.42	0.72	1.86	0.69

MT, method; B, relative median bias ×100; SD*, simulation standard deviation based on quantiles ×100; SD, simulation standard deviation ×100; SE, estimated standard error ×100; CP, 95% coverage probability based on the Wald confidence interval; MSE, mean squared error; M1, conditional logistic analysis when true X is used; M2, conditional logistic analysis when X is replaced by W; M3, two-step method; M4, proposed efficient estimator; MC1, $\alpha_0 = 0.2, \alpha_1 = 0.1$; MC2, $\alpha_0 = 0.2, \alpha_1 = 0.05$; MC3, $\alpha_0 = 0.1, \alpha_1 = 0.05$, $\text{pr}(X = 1|S, X^*) = H(-1 + S + \gamma_2 X^*)$; S1, $\gamma_2 = 1$; S2, $\gamma_2 = 2$, $\text{pr}(Y = 1|S, X, Z) = H(-2 - 2S + X + 0.5Z)$, $\alpha_0 = \text{pr}(W = 1|X = 0), \alpha_1 = \text{pr}(W = 0|X = 1)$.

Table 3. *Results of the simulation study with* $n = 1800$ *in the case of multiple instruments and multiple confounding variables*

MT	M1		M2		M3		M4	
	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
Bias	−0.03	0.34	−13.98	−0.04	−0.99	0.38	0.80	0.34
SD*	7.88	8.12	9.84	8.13	12.90	8.44	12.48	8.27
SD	7.94	8.13	9.93	8.08	13.24	8.34	12.84	8.22
SE	7.80	7.88	9.67	7.83	13.45	8.09	12.63	7.98
CP	94.25	94.47	82.46	94.55	94.18	94.50	94.60	94.64
MSE	0.63	0.66	1.93	0.65	1.75	0.70	1.66	0.68

MT	M1			M2			M3			M4		
	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3
Bias	−0.52	−0.32	1.02	−13.78	0.05	53.11	−1.17	0.06	1.94	0.20	−0.01	0.04
SD*	8.12	7.96	10.96	10.20	7.95	0.90	12.97	8.14	13.64	12.47	8.07	12.66
SD	8.03	8.07	11.13	10.16	8.03	10.93	13.10	8.22	13.71	12.61	8.16	12.85
SE	7.83	7.89	11.04	9.77	7.84	10.85	14.79	8.19	15.71	12.51	7.93	12.50
CP	97.00	94.48	94.96	82.56	94.46	83.30	94.66	94.94	94.66	94.42	94.22	93.96
MSE	0.65	0.65	1.24	1.93	0.64	2.29	1.72	0.68	1.88	1.59	0.67	1.65

Panel 1, all confounding variables are used in matching; Panel 2, all but one confounding variables are used in matching and the other confounding variable is used as a covariate for adjustment; MT, method; Bias, relative median bias ×100; SD*, simulation standard deviation based on quantiles ×100; SD, simulation standard deviation ×100; SE, estimated standard error ×100; CP, 95% coverage probability based on the Wald confidence interval; MSE, mean squared error; M1, conditional logistic analysis when true X is used; M2, Conditional logistic analysis when X is replaced by W; M3, proposed two-step method; M4, proposed efficient estimator.

in M3, while 0–0.5% data sets faced convergence issue in M4. Methods M1 and M2 never showed any convergence problem.

4.3. *Results*

Table 1 displays results for the scenario when the misclassification probabilities are the same. Naturally, the performance of M1 is the best in terms of all measures as the true values of X were used. The estimator of β_2 under different methods performs equally. On the other hand, the performance of $\hat{\beta}_1$, the coefficient for X that is not observed in the data, varies widely across the methods and scenarios. For all scenarios and both sample sizes, the bias of $\hat{\beta}_1$ under method M2 is huge, and consequently, the coverage probability for β_1 is far away from the nominal level of 0.95. The bias and variance under M3 and M4 decrease with an increasing association between X and X^*. Also, as the association between X and X^* gets stronger, the coverage probability gets closer to the nominal level. The performance of M3 is good for a large sample size, while for almost all scenarios M4 is the best in terms of reducing bias, and low MSE. Also, as expected M4 shows less variability of the estimator than M3. When $n = 1000$, we see that the bias decreases when we go from a moderate (S1) to strong (S2) association between X and X^*. However, the increased sample size greatly decreases the standard errors of our methods by nearly half in all three scenarios: MC1, MC2, and MC3. Consequently, the increase in sample size leads to a decrease in MSE, proving superiority of M3 and M4 over M2.

We now consider Table 2 that contains simulation results for unequal misclassification probabilities. For both sample sizes, M3 outperforms M2 in terms of bias for scenarios for moderate and strong association

between X and X^*, and usually for a large sample size. Under M3, as before, the bias, variability, and the distance between the coverage probability and its nominal level for the estimator of β_1 are decreasing with the increasing association between X and X^*. As in the equal misclassification probability case, moving from $n = 200$ to $n = 1000$ and increasing the association between X and X^* decreases the standard errors and MSE of our estimator greatly, giving M3 and M4 an advantage over M2. Overall, M4 beats all other approaches in terms of bias, MSE and variability.

Finally, we also considered a scenario in which we have multiple instruments and multiple confounding variables. For this purpose, we simulated data sets closely following the real data set with multiple confounding and instrumental variables; see Section 5 for more discussion on the data. We considered the subset of the data where the mother of the newborn was listed as a Black, and after necessary exclusions we ended up with $N = 42\,933$ subjects. To simulate a pseudo population we sampled 42 933 subjects with replacement from this population. Once a subject is selected in this pseudo population, values corresponding to confounding variables, covariates, and instrumental variables are automatically assigned. To generate the true exposure variable X, we use the following logistic model

$$\begin{aligned}\text{pr}(X = 1|\boldsymbol{S}, \boldsymbol{X}^*) = H(&-3.39 + 1.29S_1 + 0.36S_2 + 1.66S_3 + 0.67S_4 + 0.91S_5 + 1.09S_6 + 0.28S_7 \\ &+ 0.14X_1^* - 0.32X_2^* - 0.21X_3^* + 0.99X_4^*),\end{aligned} \tag{4.7}$$

where $S_1 = 1$ if the number of years of education of the mother < 12 and 0 otherwise, S_2, S_3 are the dummy variables corresponding second or third trimester, respectively, when the prenatal care began in the second and third trimester, S_4, S_5, S_6, S_7 are the dummy variables corresponding to the mother's age in [23, 26], [27, 30], [31, 39], ≥ 40, X_1^* = cigarette state tax rate for 1989, X_2^* denotes the logarithm of the median family income of the county where the child was born divided by 1000 minus the average of the logarithm of the median family income expressed in thousands, $X_3^* = 1$ if the father's race was Black and 0 for White, and $X_4^* = 1$ if the number of years of education of the father < 12 and 0 otherwise. The coefficients used in Model (4.7) are close to the estimates of $\boldsymbol{\gamma}$ parameters from the analysis of the data on Black mothers. Then we created the surrogate variable W as $W = B \times X$, where $B \sim \text{Bernoulli}(1 - \alpha_1)$, $\alpha_1 = 0.45$ that was the estimated α_1 in the real data analysis using method M3 under $\alpha_0 = 0$. Next we simulated the binary response Y using the following model

$$\begin{aligned}\text{pr}(Y = 1|\boldsymbol{S}, X, Z) = H(&-3.2 + 0.19S_1 - 0.03S_2 + 0.37S_3 + 0.09S_4 + 0.32S_5 + 0.57S_6 + 0.55S_7 \\ &+ 0.69X + 0.33Z),\end{aligned}$$

where the intercept and coefficients corresponding to $\boldsymbol{S}$ were obtained by regressing Y on $\boldsymbol{S}$ in the population of the Black mothers, and $Z = 1$ if the father's age was greater or equal to 31 and 0 otherwise. The coefficients corresponding to X and Z are the estimates of β_1 and β_{22} in the real data analysis with the Black mother using the proposed approach and when $\alpha_0 = 0$. This resulted in approximately 4.5% of the data with $Y = 1$. After creating this pseudo-population, we created a nested case–control data by sampling $n = 1800$ cases and 3600 matched controls. This procedure of creating a pseudo-population and sampling a nested case–control data from it was repeated 5000 times.

Each simulated nested case–control data set was analyzed by the three methods mentioned earlier, M1, M2, M3, M4. In M3 and M4 we set $\alpha_0 = 0$. The results are given in the first panel of Table 3. It is seen that the bias of estimating β_1 is hugely reduced in M3 and M4 compared with M2, and the coverage probabilities under M3 and M4 are much closer to the nominal level than that of M2. The huge bias reduction in M3 and M4 is also reflected in the reduced MSE. Specifically, compared with M2, the MSE for β_1 is 10% and 14% less in M3 and M4, respectively. Overall M4 is superior.

Within the paradigm of the last simulation design, we have considered another study where S_1, one of the confounding variables, was not used to form matched sets. However, this variable S_1 was used in

the incidence model for Y and was treated as a prognostic factor (like Z_2 according to our notation). We estimated the corresponding regression parameter, denoted by β_3, along with the regression parameters for X and the original Z. The results are presented in the second panel of Table 3. Once again, we see that both M3 and M4 perform very well with M4 being slightly better than M3.

In summary, the simulation studies indicate that the proposed methods work very well in reducing the bias significantly compared with the naive approach. It appears that the proposed methods work very well when the association between the true exposure and the instruments is strong and the sample size is large. When there is any prior information on the misclassification probabilities, incorporation of that information in the analysis would definitely help to improve the performance of the method.

5. REAL DATA ANALYSIS

5.1. *Description of the data and variables*

We consider the data from the 1989 US Natality Birth Records (National Center for Health Statistics, 1992), introduced in the Section 1. It contains information on the birth records for infants born to residents and non-residents within the United States during the year 1989. Characteristics on the mother, the father, and the child were recorded. For our analysis define $Y = 1$ if a newborn's birth weight is less than 2500 g and 0 otherwise and define $X = 1$ if a mother smoked more than 2 cigarettes daily during the pregnancy and 0 otherwise. Since there is no consensus on the definition of various levels of smoking, we took 2 as the cutoff to distinguish between (i) no, intermittent, and very light smokers, and (ii) light, moderate, and heavy smokers. The binary variable based on the reported average daily number of cigarettes smoked ($>$ 2 cigarettes as 1 and $\leq$ 2 as 0) was considered as W. Our intuition is that those who smoked more or regularly during pregnancy are more likely to have recall bias in reporting the average number than those who didn't smoke or smoked very little or infrequently. We consider the following variables as instrument: cigarette state tax rate for 1989, the logarithm of the median family income of the county where the child was born divided by 1000 minus the average of the logarithm of the median family income expressed in thousands, race of the father, and a binary variable for the number of years of education of the father ($<$ 12 and $\geq$ 12). It has been shown in the literature that cigarette tax, father's education, and family income may well serve as instrumental variables for mother's smoking (Townsend *and others*, 1994; Evans and Ringel, 1999; Martin *and others*, 2007). However, family income may have direct influence on the birthweight. Therefore, we considered median family income of the county where the mother lived as one of the instruments. Moreover, the family income was not reported in the data set, so we could not even use it as a confounding variable. We have chosen mother's age, mother's education, and when prenatal care began as confounding variables because they likely to have a direct impact on the response (birthweight) as well as on the smoking behavior of the mother. The age of the mother was split into five categories, 18–22 as the reference category, 23–26, 27–30, 31–40, and $>$ 40, a binary variable for mother's education (0 for $<$ 12 years and 1 for $\geq$ 12), and when prenatal care began with three categories, first trimester as the reference category, second trimester, and third trimester. There were a few subjects (Black: 1.00%, White: 0.30%) without any prenatal care and that category was combined with the prenatal care that began in the third trimester. For the purpose of this illustrative analysis, we assume that conditional on the true level of smoking, the reported smoking level, and all other variables are independent. While this assumption is difficult to validate without a validation dataset, it is plausible in this analysis.

From the cohort data, we excluded the subjects whose mother and father had a race other than Black and White. About 3% newborns had one of the parents neither Black nor White. This number includes the case where one of the parents is Black/White and the other parent is neither Black nor White. However, it does not include the case where both parents are neither Black nor White. We excluded the newborns whose mother or father was less than 18 years of age. Additionally, we considered only the newborns that

were the first child to his/her parents. The birth cohort was first divided into two groups: Black and White mothers, and conducted the analysis of these two groups separately. We excluded the newborns that were born in Alaska, Delaware, Montana, New Hampshire, and Oregon as these states did not have a state tax on cigarettes in 1989. In addition, we removed newborns from Colorado, Maryland, New Jersey, Rhode Island, and Wyoming because their tax rate did not apply to cigarettes. After all these exclusions, we were left with 42 933 newborns to Black mothers (group 1) and 347 041 newborns to White mothers (group 2). We conducted separate analyses for Black and White groups.

There were 2021 newborns with $Y = 1$ in group 1. We then randomly selected 2000 subjects out of 2021 subjects as cases, and for each selected subject, we randomly sampled $M = 2$ controls from 40 912 by matching the confounding variables. This resulted in 2000 cases and 4000 controls in our matched data set. There were 6646 newborns with $Y = 1$ in group 2, and following the above sampling mechanism we formed an 1:2 matched case–control data from group 2.

5.2. *Results*

We consider three methods of analysis, M2 (naive), M3 (intuitive two stage method), and M4 (efficient method). Since the true X is never observed in this real data, method M1 cannot be applied. Results for groups 1 and 2 are presented in the first and second panels, respectively, of Table 4. We performed two analyses under M3 and M4. Initially, we discuss the results for group 1. First, using the regular identifiability condition $0 < \alpha_0 + \alpha_1 < 1$, we found that estimated α_0 and α_1 were 0.014 (s.e. 0.013) and 0.521 (s.e. 0.148), respectively in M3. In the real data analyses, the estimates of α_0 and α_1 in M3 and M4 are quite close, so we present these estimates based on M3 only. Based on the above estimates and their standard errors, we concluded that α_0 was not significantly different from zero. Therefore, we reanalyzed the data by setting $\alpha_0 = 0$, and estimated all the parameters along with $\alpha_1 \in (0, 1)$. In this scenario α_1 was estimated as 0.46 (s.e. 0.112). All three methods M2, M3, M4 indicate that mother's smoking is positively associated with low birthweight. Based on the second analysis where $\alpha_0 = 0$, the estimated odds ratios of smoking are 2.15 and 1.99 for M3 and M4, while that odds ratio estimate based on M2 is 1.7. That means we observe approximately 26% and 17% increase in the odds ratio estimate from M2 to M3 and M2 to M4. Father's age shows statistically significant association with the risk of low birthweight for all scenarios except for M3 when $\alpha_0 = 0$. The standard error for β_1 is slightly larger in M3 and M4 than in M2. This increase is natural as in M3 and M4 we take into account the uncertainty of not observing the true values of the exposure.

For group 2, we found that under the regular identifiability condition $0 < \alpha_0 + \alpha_1 < 1$, estimates for α_0 and α_1 were 0.014 (s.e. 0.003) and 0.271 (s.e. 0.096), respectively, for M3. We also conducted the second analysis by setting $\alpha_0 = 0$. In the second analysis, we obtained $\hat{\alpha}_1 = 0.251$ (s.e. 0.089) for M3. Similar to the black mothers, M2, M3, M4 indicate that white mother's smoking is positively associated with low birthweight. However, it appears that both categories of father's age have statistically significant association with the low birthweight in all methods and for both cases where $\alpha_0 = 0$ and $\alpha_0 > 0$.

6. Discussion

In this article, we proposed two consistent methods for reducing the bias when estimating the association parameters in a matched case–control study. The novelty of the methods is to make use of instruments to recover the measurement uncertainty in the absence of validation data. The methodology is accompanied with an uncertainty measure, and contains a theoretical justification of the large sample properties. The realistic simulation studies shed lights on the performance of the proposed methods. Specifically, the simulation results indicate that the performance of the proposed methods is satisfactory for a large sample size, strong association between the instruments and the true exposure, and moderate misclassification

Table 4. Analysis of the low birthweight data sampled from 1989 US birth cohort

MT		M2			M3: $0 < \alpha_0 + \alpha_1 < 1$			M3: $\alpha_0 = 0, 0 < \alpha_1 < 1$			M4: $0 < \alpha_0 + \alpha_1 < 1$			M4: $\alpha_0 = 0, 0 < \alpha_1 < 1$		
Mother		EST	SE	PV	EST	SE	PV	EST	SE	PV	EST	SE	PV	EST	SE	PV
Black	β_1	0.532	0.097	0.000	0.824	0.170	0.000	0.764	0.130	0.000	0.613	0.117	0.000	0.685	0.141	0.000
	β_{21}	−0.096	0.067	0.149	−0.119	0.075	0.114	−0.139	0.079	0.078	−0.088	0.078	0.263	−0.127	0.077	0.099
	β_{22}	0.362	0.133	0.007	0.309	0.144	0.031	0.218	0.161	0.177	0.425	0.129	0.001	0.288	0.129	0.025
White	β_1	0.396	0.072	0.000	0.451	0.092	0.000	0.460	0.130	0.000	0.488	0.073	0.000	0.468	0.084	0.000
	β_{21}	0.420	0.065	0.000	0.424	0.072	0.000	0.427	0.072	0.000	0.414	0.073	0.000	0.429	0.065	0.000
	β_{22}	0.659	0.143	0.000	0.667	0.145	0.000	0.683	0.149	0.000	0.652	0.151	0.000	0.667	0.144	0.000

Here, β_1 is the regression coefficient for mother's smoking, and β_{21} and β_{22} are the regression coefficients for father's age in $[31, 40]$ and > 40, respectively, while the father's age in $[18, 30]$ is treated as the reference category. MT, method; M2, naive approach; M3, proposed two-step method; M4, proposed efficient estimator; EST, estimate; SE, standard error; PV, p-value.

probabilities. We believe that this is the first work that makes use of instruments to reduce misclassification bias in the absence of any prior knowledge on the misclassification probability or a validation data set. The basic idea may be generalized to a multicategory exposure variable with added complexity of estimating comparatively a large number of misclassification probabilities. The proposed methodology can also be used when the instrumental variables are observed for a subset of the main data set. Following a referee's comment we want to point out that in our illustrative data example the nondifferential misclassification assumption may be violated due to dichotomization of the numeric exposure variable (Flegal *and others*, 1991), and developing a proper methodology for dealing with this scenario is a part of our future research. To handle potential convergence problems in the proposed methods one may consider a penalized estimator using Firth (1993)'s penalty function.

7. Software

The code for computation for the scenario of Tables 1, 2, and 3 is available at https://github.com/ronsami/Istrumental-variables-for-misclassification.

Supplementary material

Supplementary material is available at http://biostatistics.oxfordjournals.org.

Acknowledgments

The authors wish to thank the reviewer and the associate editor for their valuable comments and suggestions which substantially improved an earlier version of the paper. *Conflict of Interest*: None declared.

Funding

This research was supported in part by the Simons Foundation Mathematics and Physical Sciences—Collaboration Grants for Mathematicians Program Award 499650.

References

Bound, J. and Krueger, A. (1991). The extent of measurement error in longitudinal earnings data: do two wrongs make a right. *Journal of Labor Economics* **12**, 1–24.

Chu, R., Gustafson, P. and Le, N. (2010). Bayesian adjustment for exposure misclassification in case-control studies. *Statistics in Medicine* **29**, 994–1003.

Duffy, S. W., Rohan, T. E., Kandel, R., Prevost, T. C., Rice, K. and Myles, J. P. (2003). Misclassification in a matched case-control study with variable matching ratio: application to a study of c-erbb-2 overexpression and breast cancer. *Statistics in Medicine* **22**, 2459–2468.

Evans, W. N. and Ringel, J. S. (1999). Can higher cigarette taxes improve birth outcomes. *Journal of Public Economics* **72**, 135–154.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika* **80**, 27–38.

Flegal, K., Keyl, P. and Nieto, F. (1991). Differential misclassification arising from nondifferential errors in exposure measurement. *American Journal of Epidemiology* **134**, 1233–1244.

Godambe, V. (1976). Conditional likelihood and unconditional optimum estimating equations. *Biometrika* **63**, 277–284.

Greenland, S. (2000). An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology* **29**, 722–729.

GREENLAND, S., ROBINS, J. M. AND PEARL, J. (1999). Confounding and collapsibility in causal inference. *Statistical Science* **14**, 29–46.

HAUSMAN, J. W., NEWEY, W., ICHIMURA, H. AND POWELL, J. (1991). Measurement errors in polynomial regression models. *Journal of Econometrics* **50**, 273–295.

HAUSMAN, J. A., ABREVAYA, J. AND SCOTT-MORTON, F. M. (1998). Misclassification of the dependent variable in a discrete-response setting. *Journal of Econometrics* **87**, 239–269.

HERNÁN, M. A. AND ROBINS, J. M. (2008). *Causal Inference*. Boca Raton: Chapman and Hall, CRC.

HU, Y. (2008). Identification and estimation of nonlinear models with misclassification error using instrumental variables: a general solution. *Journal of Econometrics* **144**, 27–61.

LIU, J., GUSTAFSON, P., CHERRY, N. AND BURSTYN, I. (2009). Bayesian analysis of a matched case-control study with expert prior information on both the misclassification of exposure and the exposure-disease association. *Statistics in Medicine* **28**, 3411–3423.

LYLES, R. H., ZHANG, F. AND DREWS-BOTSCH, C. (2007). Combining internal and external validation data to correct for exposure misclassification: a case study. *Epidemiology* **18**, 321–328.

MARTIN, L., MCNAMARA, M., MILOT, A. S., HALLE, T. AND HAIR, E. (2007). The effects of father involvement during pregnancy on receipt of prenatal care and maternal smoking. *Maternal and Child Health Journal* **11**, 595–602.

MORRISSEY, M. J. AND SPIEGELMAN, D. (1999). Matrix methods for estimating odds ratios with misclassified exposure data: extensions and comparisons. *Biometrics* **55**, 338–344.

NATIONAL CENTER FOR HEALTH STATISTICS. (1992). *Data File Documentations, Natality, 1989*. Data retrieved from NCHS' Vital Statistics Natality Birth Data, http://www.nber.org/data/vital-statistics-natality-data.html.

PRESCOTT, G. J. AND GARTHWAITE, P. H. (2005). Bayesian analysis of misclassified binary data from a matched casecontrol study with a validation sub-study. *Statistics in Medicine* **24**, 379–401.

RATHOUZ, P., SATTEN, G. AND CARROLL, R. (2002). Semiparametric inference in matched case-control studies with missing covariate data. *Biometrika* **89**, 905–916.

RICE, K. (2003). Full-likelihood approaches to misclassification of a binary exposure in matched case-control studies. *Statistics in Medicine* **22**, 3177–3194.

SCHENNACH, S. (2007). Instrumental variable estimation of nonlinear errors-in-variables models. *Econometrica* **75**, 201–239.

TOWNSEND, J., RODERICK, P. AND COOPER, J. (1994). Cigarette smoking by socioeconomic group, sex, and age: effects of price, income, and health publicity. *BMJ* **309**, 923–927.

[*Received July 20, 2018; revised January 13, 2019; accepted for publication April 2, 2019*]