Bias Analysis for Misclassification Errors in both the Response Variable and Covariate

Juxin Liu\textsuperscript{a}, Annshirley Afful\textsuperscript{a}, Holly Mansell\textsuperscript{b}, and Yanyuan Ma\textsuperscript{c}

\textsuperscript{a}Department of Mathematics and Statistics, University of Saskatchewan, Saskatoon, Canada; \textsuperscript{b}College of Pharmacy and Nutrition, University of Saskatchewan, Saskatoon, Canada; \textsuperscript{c}Department of Statistics, Pennsylvania State University, University Park, PA

ABSTRACT

Much literature has focused on statistical inference for misclassified response variables or misclassified covariates. However, misclassification in both the response variable and the covariate has received very limited attention within applied fields and the statistics community. In situations where the response variable and the covariate are simultaneously subject to misclassification errors, an assumption of independent misclassification is often used for convenience without justification. This article aims to show the harmful consequences of inappropriate adjustment for joint misclassification errors. In particular, we focus on the wrong adjustment by ignoring the dependence between the misclassification process of the response variable and the covariate. In this article, the dependence of misclassification in both variables is characterized by covariance-type parameters. We extend the original definition of dependence parameters to a more general setting. We discover a single quantity that governs the dependence of the two misclassification processes. Moreover, we propose likelihood ratio tests to check the nondifferential/independent misclassification assumption in main study/internal validation study designs. Our simulation studies indicate that ignoring the dependent error structure can be even worse than ignoring all the misclassification errors when the validation data size is relatively small. The methodology is illustrated by a real data example.

1. Introduction

Categorical data subject to misclassification errors are commonly encountered in many applied fields. In a self-report questionnaire, for example, participants may fail to report their true behavior regarding sensitive topics (e.g., self-reported alcohol or substance use during pregnancy). It has long been recognized that simply ignoring misclassification can lead to biased inference that may severely distort the truth. As such, much literature has focused on methods to address misclassification error in a single variable (either a covariate or a response variable). To name a few, Carroll et al. (2006), presented a systematic discussion, including nonlinear regression models, on this topic. Gustafson (2004) provided a textbook treatment of Bayesian adjustments for misclassification with applications in epidemiology. Yi (2017) summarized the most updated overview of statistical analysis with misclassification that covers some complex modeling settings (e.g., longitudinal data and survival data).

However, literature studying simultaneous misclassification errors in both the response variable and the covariate is rather scant even though such simultaneous errors are commonly encountered (van Smeden, Lash, and Groenwold 2020). For instance, patient reported outcome measures in the form of self-report survey data have become increasingly common endpoints of clinical studies. The information on both the response variable (such as birth weights of newborns) and the covariates (such as alcohol use or substance use during pregnancy) may be subject to some shared sources of error. Methodological work is therefore needed to handle this type of misclassification in both variables.

An early work by Barron (1977) aimed to uncover the true relative risk after adjusting for the independent misclassification in the response variable and the covariate (both binary), assuming reasonable estimates of misclassification parameters (sensitivity and specificity) are available. Brenner, Savitz, and Gefeller (1993) and Vogel et al. (2005) extended the work to allow dependent misclassification errors and the dependence is characterized by a set of covariance-type dependence parameters. Different from the notion of the covariance-type of dependence parameters, Tang et al. (2013) and Salway et al. (2019) considered factorizing the observed data joint conditional distribution to account for misclassification dependence. The aforementioned literature on simultaneous misclassification errors handled only binary variables. Greenland and Kleinbaum (1983) briefly mentioned a general form for categorical response and covariates, but restricted it to independent misclassification. To the best of our knowledge, we are the first to study the consequences of wrong adjustment for the simultaneous misclassification errors in both the response variable and covariate.

The novel contributions of our work are outlined below. First, we extend the definition of dependence parameters in Vogel et al. (2005) to the more general setting of differential...
misclassification errors. That is, response variables (covariates) provide additional information about how covariates (response variables) are missclassified. Second, we propose the use of likelihood ratio tests to verify assumptions of nondifferential and/or independent misclassification errors. Third, we conduct simulation studies to assess the impact of ignoring the dependence between the two misclassification processes.

The remainder of this article is outlined as follows. Section 2 presents the data structure, model frame, and notations for both binary variables and categorial variables. Section 3 explores the asymptotic bias in odds ratio when ignoring all the misclassification error issues (no adjustment at all). To investigate the consequences of ignoring the dependence in the simultaneous misclassification mechanism, Section 4 presents the simulation study and the results for both binary variables and categorical variables. A real data example is illustrated in Section 5. Section 6 concludes the findings and discusses possible future work.

2. Data, Modeling, and Notation

2.1. Data Structure

In this article, we consider main study/internal validation study designs, which are often used in practice if variables of interest are expensive or difficult to measure. For example, liver biopsy is the gold standard for measuring liver fibrosis but it is invasive and expensive. Therefore, in some circumstances it may only be available for a small proportion of the participants in the sample. Transient elastography (Fibroscan) is a noninvasive way to measure liver fibrosis and is being increasingly used to diagnose liver disease. Such type of surrogate measurement of fibrosis may be more readily available and in this example is applied to all participants in the study. That is, for a random subsample \((i = 1, \ldots, n_v)\), one can observe \((Y^*, X^*, Y, X)\). For the remaining sample \((i = n_v + 1, \ldots, n)\) only \((Y^*, X^*)\) can be observed. Therefore, the validation data is in the form of \((Y^*_i, X^*_i, Y_i, X_i)\) for \(i = 1, 2, \ldots, n_v\), while the main data is in the form of \((Y_i, X_i)\) for \(i = n_v + 1, \ldots, n\).

2.2. Dependence Parameters

The data structure considered in our work is based on main study/internal validation study designs. The reason is 2-fold. First, main study/internal validation study designs are a popular choice when it is time and cost expensive to collect error-free data. Second, it is theoretically challenging to consider the data structure with contaminated data only. The challenge is model identification; it is too ambitious to infer such a rich dependent misclassification error model from contaminated data alone.

Let \(Y^*\) represent the observed error-contaminated outcome variable and \(X^*\) the observed error-contaminated covariate. Let \(Y\) and \(X\) denote the underlying true variables, respectively. The joint misclassification modeling proposed in our work can be easily adapted to situations where an error-free categorical covariate, say \(Z\), is available. That is, \((Y^*, X^*)/Y, X, Z)\) can be built upon the proposed modeling for each data stratum formed by \(Z\).

Following the rationale behind the definition in Brenner, Savitz, and Gefeller (1993) and Vogel et al. (2005), the dependence parameters between the misclassification process for \(Y\) and that for \(X\) are defined as

\[
D_{yx:yx} = P(Y^* = y^*, X^* = x^*|Y = y, X = x) - P(Y^* = y^*|Y = y, X = x)P(X^* = x^*|Y = y, X = x),
\]

where \(y = 0, 1, \ldots, I - 1, x = 0, 1, \ldots, J - 1\). Thus, there are a total of \(IJ\) different dependent parameters. Throughout the article, we use uppercase letters for random variables and lowercase letters for their realized values.

At first glance, one can see that each dependence parameter is essentially conditional covariance and thus takes values between \(-1\) and \(1\). We give more refined bounds in supplementary note 4.

The above formulation (1) is for most general setting with minimal assumption for the joint misclassification errors in both variables. It includes some commonly seen models as special cases. First, if all the \(D\) parameters are zero, we have independent misclassification models:

\[
P(Y^* = y^*, X^* = x^*|Y = y, X = x) = P(Y^* = y^*|Y = y, X = x)P(X^* = x^*|Y = y, X = x).
\]

Note the nondifferential errors are included as a special case as \(X\) can be dropped from \(P(Y^*|Y, X)\) and/or \(Y\) can be dropped from \(P(X^*|X, Y)\).

Second, if both variables are subject to nondifferential misclassification errors, we have dependent models for nondifferential errors in both variables. This is equivalent to the model Vogel et al. (2005) considered.

Third, if nondifferential misclassification errors occur in one variable and differential errors in the other variable, we can have the dependent models in the same notion as (1). Suppose \(Y\) is subject to nondifferential errors and \(X\) is subject to differential errors,

\[
D_{yx:yx} = P(Y^* = y^*, X^* = x^*|Y = y, X = x) - P(Y^* = y^*|Y = y)P(X^* = x^*|Y = y, X = x).
\]

Thanks to the fact that the aforementioned three cases are nested within the most general form (1), we propose likelihood ratio tests for testing nondifferential and/or independent assumptions.

In the following, we will first present a simple case where both \(X\) and \(Y\) are binary, then extend the discussion to categorical variables.

2.3. Binary \(Y\) and \(X\)

When all variables \(Y, X, Y^*, X^*\) are binary, there is some simplification in the \(D\) parameters. We point out that we only need to define the dependence parameter through considering the same value \(y\) for the \((Y, Y^*)\) pair and the same value \(x\) for the \((X, X^*)\) pair. As such, the notations for \(D\) simplify as follows:

\[
D_{yx} = P(Y^* = y, X^* = x|Y = y, X = x) - P(Y^* = y|Y = y, X = x)P(X^* = x|Y = y, X = x).
\]
Note that the definition reduces to the dependence parameters defined by Vogel et al. (2005) under nondifferential error assumption. That is, the covariate (response variable) does not provide any additional information about how the response variable (covariate) is misclassified. In total, there are 2^2 free \( D \) parameters instead of 2^4 \( D \) parameters due to the relations proved in supplementary note 1.

It is worth noting that the four \( Dyx \) dependence parameters are closely related to conditional covariance, that is, differential and dependent misclassification are closely related to conditional covariance, that is,

\[
pyx = \text{cov}(Y^*, X^* | Y = y, X = x),
\]

where \( I(\cdot) \) is the indicator function valued 1 if the condition holds and 0 otherwise.

The misclassification parameters, that is, sensitivity and specificity, are defined as follows:

\[
SN_{YX} = P(Y^* = 1 | Y = 1, X = x), \\
SP_{YX} = P(Y^* = 0 | Y = 0, X = x), \\
SN_{XY} = P(X^* = 1 | X = 1, Y = y), \\
SP_{XY} = P(X^* = 0 | X = 0, Y = y).
\]

Let \( \mathbf{p} = (p_{11}, p_{10}, p_{01}, p_{00})' \) and \( \mathbf{p}^* = (p_{11}^*, p_{10}^*, p_{01}^*, p_{00}^*)' \), where \( p_{yx} = P(Y = y, X = x) \) and \( p_{yx}^* = P(Y = y, X = x) \).

Thanks to the total probability rule, the relationship between \( \mathbf{p} \) and \( \mathbf{p}^* \) can be expressed in the matrix form \( \mathbf{p}^* = \mathbf{M} \mathbf{p} \) (Please see supplementary note 2 for the details of \( \mathbf{M} \)).

Let \( \mathbf{\theta} \) represent the collection of all the parameters; that is,

\[
\mathbf{\theta} = (SN_{YX}, SP_{YX}, SN_{XY}, SP_{XY}, p_{11}, p_{10}, p_{01}, D_{11}, D_{10}, D_{01}, D_{00})',
\]

where \( x, y = 0, 1 \). Note \( p_{00} \) is omitted due to the fact \( p_{00} = 1 - p_{10} - p_{01} - p_{11} \).

Let \( L_\nu(\mathbf{\theta}) \) represent the likelihood function for the validation data and \( L_m(\mathbf{\theta}) \) for the main data. Under the most general scenario, that is, differential and dependent misclassification errors, one can easily derive

\[
L_\nu(\mathbf{\theta}) = \prod_{i=1}^n P(Y_i^* = y_i^*, X_i^* = x_i^*, Y_i = y_i, X_i = x_i)
\]

\[
= \prod_{i=1}^n P(Y_i^* = y_i^*, X_i^* = x_i^* | Y_i = y_i, X_i = x_i)
\]

\[
P(Y = y, X = x).
\]

\[
L_m(\mathbf{\theta}) = \prod_{i=n_0+1}^n P(Y_i^* = y_i, X_i^* = x_i)
\]

\[
= \prod_{i=n_0+1}^n \sum_{yx} P(Y_i^* = y_i, X_i^* = x_i | Y = y, X = x)
\]

Please note that the derivation of \( L_m \) can be simplified under the assumption of independent misclassification errors, that is,

\[
L_{m,\text{ind}}(\mathbf{\theta}) = \prod_{i=n_0+1}^n \sum_{yx} P(Y_i^* = y_i | Y = y)P(X_i^* = x_i | X = x)
\]

\[
P(Y = y, X = x).
\]

As shown in supplementary note 3,

\[
L_m(\mathbf{\theta}) = \prod_{i=n_0+1}^n (L_{m,\text{ind}}(\mathbf{\theta} | y_i^*, x_i^*) + (-1)^{y_i^*+x_i^*}\delta),
\]

where

\[
\delta = \sum_{yx} (-1)^{y+x} D_{yx} p_{yx},
\]

\[
L_{m,\text{ind}}(\mathbf{\theta} | y_i^*, x_i^*) = \sum_{yx} p_{yx} \left[ \left( SN_{YX}^* \right)^{1-y} \left( 1 - SN_{YX} \right)^{1-y} \right]^{y-x}
\]

\[
\left[ \left( 1 - SP_{XY} \right)^{y} \left( SP_{XY} \right)^{1-y} \right]^{1-x}.\]

Clearly one can see that the difference between \( L_m \) and \( L_{m,\text{ind}} \) relies on \( \delta \) only.

We show that \( \delta = E(\text{cov}(Y^*, X^* | Y, X)) \) in supplementary note 5, which provides an intuitive interpretation for \( \delta \). A scaled version of the dependence measure, denoted by \( \delta_r \), is given by the following:

\[
\delta_r = E \text{ corr}(Y^*, X^* | Y, X).
\]

Because the validation data is independent of the main data, the overall likelihood function, denoted by \( L(\cdot) \), can be derived as

\[
L(\mathbf{\theta}) = L_\nu(\mathbf{\theta}) L_m(\mathbf{\theta}).
\]

There is no closed-form of the MLE estimators by maximizing the \( L(\mathbf{\theta}) \), mainly due to the nonlinear constraints that are intrinsic to the definitions of \( D_{yx} \) plus additional constraints on misclassification parameters. We consider a Bayesian approach with very diffuse priors, which leads to the posterior estimates very close to the MLE. The Bayesian MCMC sampling is implemented by an R package “runjags.”

### 2.4. Categorical Y and/or X

In this section, we provide details for categorical variables on the constraints for the \( D \) parameters and the single quantity that can be used to evaluate the dependence between the two misclassification processes. The section is more general and includes binary variables as a special case. Moreover, it also includes a mixed case of one binary variable and one categorical variable as a special case. Please be noted that the likelihood function derivation in main study/internal validation study designs is similar to that for binary variables in Section 2.2 and thus omitted here.

Thanks to the fact that

\[
\sum_y D_{yx, st} = 0,
\]

\[
\sum_x D_{yx, st} = 0,
\]

the number of free \( D \) parameters is \((I-1)(J-1)IJ\). Though the number of \( D \) parameters increases in the order of \( I^2J^2\),
the discussion for categorical variables still has important value in applications where the category number of the misclassified variable is 3 or 4. For example, infectious status can be categorized as no infection, mild infection, and severe infection. As for categorical variables with more than four levels, we would suggest treating such variables as continuous if merging some categories is not an option.

Now let \( p \) be the corresponding vector with elements of \( p_{yx} \) in order of \( (0, 0, 1, 0, 2, 0, 1) \) and merging some categories is not an option. For categorical variables with more than four levels, we would suggest treating such variables as continuous if merging some categories is not an option.

For categorical variables with more than four levels, we would suggest treating such variables as continuous if merging some categories is not an option.

### 3. Bias in Odds Ratio When Misspecifying Misclassification Models

Previous literature (e.g., Kristensen 1992; Vogel et al. 2005) considers the effects of misclassification errors in bias analysis when completely ignoring misclassification errors. In this section, we consider the asymptotic bias in odds ratio (OR) in two scenarios: (i) completely ignoring misclassification issues; (ii) ignoring the dependence of the two misclassification processes. Recall that \( \mathbf{p} = (p_{11}, p_{10}, p_{01}, p_{00})' \) and \( \mathbf{p}^* = (p_{11}^*, p_{10}^*, p_{01}^*, p_{00}^*)' \). Let us first look at Scenario (i). The true OR is calculated by

\[
\text{OR} = \frac{p_{11}p_{00}}{p_{10}p_{01}}.
\]

The **apparent** odds ratio based on the error-prone variables, denoted by \( \text{OR}^* \), is similarly derived as:

\[
\text{OR}^* = \frac{p_{11}^*p_{00}^*}{p_{10}^*p_{01}^*}.
\]

Thanks to (S.1), we can write the bias in terms of OR as

\[
\log(\text{OR}^*) - \log(\text{OR}) = \log\left(\frac{p_{11}^*p_{00}^*}{p_{10}^*p_{01}^*}\right) - \log\left(\frac{p_{11}p_{00}}{p_{10}p_{01}}\right)
= g(p_0, \beta_1, p_x, S_{N_Y}, S_{N_X}, S_{P_Y}, S_{P_X}, S_{X}, S_{Y}, S_{X}, S_{Y}).
\]

Note that the dependence parameters are absorbed in \( \delta \). This is different from the bias analysis in Kristensen (1992) in two aspects: (a) The derivation in Kristensen (1992) was only for nondifferential misclassification. (b) The bias function derived in Kristensen (1992) depends on all 12 simultaneous misclassification probabilities \( P(Y^*, X^*|Y, X) \). In contrast, our function only has nine parameters for the simultaneous misclassification model. For differential misclassification, it is hard to visualize how the bias changes along with the misclassification parameters and the dependence. But for nondifferential misclassification, if one assumes \( S_{N_Y} = S_{N_X} = S_{N} \) and \( S_{P_Y} = S_{P_X} = S_{P} \)
and fix the values of $\beta_0, \beta_1, \beta_2$, the above function only depends on SN, SP, and $\delta$. Figure 1 presents the contour plot of the bias in OR when $\delta = 0$ and $\delta = 0.01$, respectively. For $\delta = 0$, the magnitude of bias in OR becomes smaller when the amount of misclassification decreases; that is, both SN and SP increase. While for $\delta = 0.01$, there is no such trend. The bias in OR decreases first, but then increases when the amount of misclassification decreases. Even when both SN and SP are close to 1, the bias in OR can be substantial.

Now let us consider Scenario (ii). To the best of our knowledge, there is no existing literature on bias analysis when misclassification errors are adjusted in a wrong way. Let $h(p_1, p_2, p_3) = p_1(1 - p_2 - p_3)/p_2p_3, \mathbf{p}^* = (p_{11}, p_{10}, p_{01})'$ and $\mathbf{p}^* = (p_{11}, p_{10}, p_{01})'$.

Using the first order Taylor expansion, we have the following

$$h(\mathbf{p}^*) \approx h(\mathbf{p}) + \nabla h(\mathbf{p})(\mathbf{p}^* - \mathbf{p}).$$

Similarly we can expand $h(\mathbf{p}^*_{\text{ind}})$ at $\mathbf{p}$. Then the difference $h(\mathbf{p}^*) - h(\mathbf{p}^*_{\text{ind}})$ is the bias caused by ignoring the dependence in misclassification errors. Let $\mathbf{p}^*_{\text{ind}}$ denote the joint distribution for $(Y^*, X^*)$ under the independent misclassification model; that is, all $D$ parameters are valued 0. Noting the fact $\mathbf{p}^* = \mathbf{p}^*_{\text{ind}} + \delta(1, -1, -1, 1)'$, immediate fact from (S.2), we have

$$h(\mathbf{p}^*) - h(\mathbf{p}^*_{\text{ind}}) \approx \nabla h(\mathbf{p})\delta(1, -1, -1, 1)' = \frac{\delta}{P_{10}P_{01}} \left[ p_{10}p_{11} + p_{11}p_{00} + p_{11}p_{00} - p_{10}p_{01} \right].$$

In case of imperfect gold standard tools, the derived formula is still valid, but no direct information on $\delta$ and $\mathbf{p}$ can be inferred from the observed data. If there is background knowledge on the accuracy of the gold standard measurement, one may consider Bayesian methods. Otherwise, one will need to consider different study designs (e.g., replicates of surrogate variables) so as to make inference for the joint misclassification model.

In Section 4, we will present simulation experiments to study the consequences of fitting an independent misclassification model to data generated from dependent misclassification model. Both binary variables and categorical variables are considered in our simulation studies.

### 4. Simulation Studies

For readers who are interested in applying the joint misclassification models mentioned above, the R code for simulation studies is publicly available at Github [https://github.com/JuxinLiu/Misclassification.git](https://github.com/JuxinLiu/Misclassification.git).

#### 4.1. Binary Variables

We conduct a numerical experiment varying the proportion size of validation data, level of misclassification (sensitivities and specificities), type of misclassification (differential or nondifferential), and $\delta$ for binary $Y$ and $X$. Here is the list of different scenarios.

- Three different choices of $n_v/n$, that is, proportion size of internal validation data: 10%, 30%, and 50%
- Four different types of simultaneous misclassification: both $Y$ and $X$ subject to nondifferential/differential errors, $Y$ subject to nondifferential misclassification and $X$ subject to differential misclassification, $Y$ subject to differential misclassification and $X$ subject to nondifferential misclassification.

For each one of the four types of misclassifications, we consider

- three different choices of SN&SP: largest, medium, and smallest values (Table 1 lists the exact values).
- two different choice of $D$ parameters: one set of $D$ to minimize $\delta_x$ and the other set to maximize $\delta_x$ when all
parameters are chosen to minimize/maximize model(all are fitted, that is, dependent error model, independent error from the above model. For each dataset, three different models we compare ignoresthemisclassificationerrors.

Interestingly, the results are quite different in the case of differential misclassification errors. As shown in Figures S.7–S.9, supplementary materials, the performances of independent and dependent models are almost identical. It seems that differential error models can help mitigate the bias caused by ignoring the dependence between the two misclassification processes. One possible reason is that differential misclassification mechanisms may have explained the dependence to some extent. However, a general conclusion must be based on theoretical investigation and is beyond the scope of this article.

In summary, deliberations are needed when constructing the misclassification model especially when the proportion of validation data is relatively small. As suggested by the findings, we recommend testing nondifferential error assumption first, as outlined in Section 2.4. The assumption of nondifferential misclassification error should be used with extra caution when both Y and X are potentially subject to the same source of errors.

### 4.2. Categorical Variables

In this section, we consider a case where both Y and X are trinary and ordinal. Let 0, 1, 2 represent the three categories with ordering nature for each variable. Even though it is straightforward to extend from binary to trinary variables conceptually, there are technical and computational challenges. The number of free D parameters is 36 and the remaining 45 D parameters are functions of the free D parameters. Recall that each of the 81 D parameters are bounded. As shown in supplementary note 4, each bound is a nonlinear function of the misclassification parameters. Due to such challenges, here we consider nondifferential misclassification only and assume misclassification errors only happen between two adjacent categories for each variable. That is, the misclassification matrices are in the following form:

\[
\begin{align*}
M_Y &= \begin{bmatrix}
C_Y & 1 - C_Y & 0 \\
1 - C_Y & C_Y & 1 - C_Y \\
0 & 1 - C_Y & C_Y
\end{bmatrix}, \\
M_X &= \begin{bmatrix}
C_X & 1 - C_X & 0 \\
1 - C_X & C_X & 1 - C_X \\
0 & 1 - C_X & C_X
\end{bmatrix},
\end{align*}
\]

where

\[
\begin{align*}
C_Y &\triangleq P(Y^* = i|Y = i), \\
C_X &\triangleq P(X^* = i|X = i), \quad i = 0, 1, 2.
\end{align*}
\]

In this setting, there are 49 D parameters in total and 16 of them are free D parameters. We again used Bayesian MCMC sampling, implemented by runjags, for the model estimation. Because of the restriction of the syntax in jags, we cannot
impose all 49 intervals for the $D$ parameters. Instead, we derived the interval bounds for each of the 16 free $D$ parameters and included them in the model file for running the jags. The remaining constraints on the functions of the free $D$ parameters are taken into consideration after the sampling is done. Only those posterior samples that satisfy all the remaining constraints are included in the posterior summary to form the point estimates.

For three different settings of $(C_X, C_Y)$ (high and low), two different choices of $\phi_c$, and three different settings of proportion size of validation data, 10%, 30%, and 50%, we consider a total of 18 different scenarios.

For each scenario, we repeatedly generate 1000 datasets from the following models:

\[
\begin{align*}
\logit P(Y \leq i|X = j) &= \alpha_i + \beta_j I\{X = j\}, i = 0, 1, j = 1, 2, \\
P(X = j) &= 1/3, \\
\alpha_0 &= -1, \alpha_1 = 1, \beta_1 = 1, \beta_2 = 2.
\end{align*}
\]

The findings are similar to that for binary variables reported in Section 4.1. All plots are included in supplementary materials. As shown by Figure S.10, supplementary materials, ignoring the dependence introduces larger bias in $\beta_1$ and $\alpha_1$ than a naïve model when the proportion of validation data is small (10%), the amount of misclassification error is largest, and the dependence is stronger. Such a phenomenon disappears when the proportion of validation data increases.

### 5. Real Data Example

We consider a real data example analyzed in Tang et al. (2013) and perform analysis to infer whether the misclassification errors in the two variables are dependent to each other. This example involves two binary variables: Trichomoniasis (TRICH) status ($Y$) and Bacterial vaginosis (BV) status ($X$) of women enrolled in the HIV Epidemiology Research Study (HERS) in four U.S. cities from 1993 to 1995 Smith et al. (1997).

Each binary variable can be assessed by two different clinical methods, one is less expensive and less accurate while the other is considered as gold standard method. There are 916 women (at the fourth HERS visit) considered in the data. For 229 of them, we use data collected from both clinical methods for assessing BV/TRICH. For the remaining 687 patients, we only use data collected from the less accurate clinical methods for BV/TRICH. For the detailed data, please refer to Table 7 in Tang et al. (2013).

Please note that the key difference between our models and that in Tang et al. (2013) is how to characterize the dependent misclassification in both the response variable and the covariate. We use the covariance like dependence parameters $D_{yx}$ while Tang et al. (2013) considers the different misclassification parameters ($SN$ and $SP$) for different strata formed by the variables that are associated with the misclassification errors. For example, to reflect the fact that misclassification errors in $Y$ are associated with $X$ and $X^*$, they consider

\[
P(Y^* = 1|Y = 1, X = x, X^* = x^*) \triangleq SN_{Yxx^*}.
\]
To be clear, the parameterization in terms of dependence parameters is not merely an equivalent parameterization to that in Tang et al. (2013). In our model setting, a single quantity $\delta$ or $\delta_r$ can be used to summarize the dependence strength between the misclassification of $Y$ and $X$. Moreover, the difference between the likelihood functions (based on main data) under the independent and dependent misclassification models lies in $\delta$ as indicated by (3). In contrast, one cannot easily tell how the two misclassification processes are correlated directly from the parameterization in Tang et al. (2013). Therefore, the parameterization here is more appealing in terms of interpretability.

We first check whether the misclassification is differential or nondifferential. Recall that

$$\theta = (SN_{YX}, SP_{YX}, SN_{XY}, SP_{XY}, p_{11}, p_{10}, p_{01}, D_{11}, D_{10}, D_{01}, D_{00})'. $$

We test $H_0 : SN_{Y1} = SN_{Y0}, SP_{Y1} = SP_{Y0}, SN_{X1} = SN_{X0}, SP_{X1} = SP_{X0}$ versus $H_1 : \text{not } H_0$. We consider a likelihood ratio test by treating the model under the null hypothesis as a special case of the model under the alternative hypothesis. Our analysis leads to

$$LRT = -2 (\log L(\hat{\theta}_0) - \log L(\hat{\theta})) = 9.49.$$ 

The $p$-value for this test is $1.6e^{-07}$. This suggests strong evidence against the null hypothesis. We further check whether misclassification in $Y$ is nondifferential. The $p$-value is extremely small and thus a strong evidence in favor of differential errors in $Y$. Finally we check whether the misclassification in $X$ is nondifferential. The $p$-value is 0.017 and we take it as significant evidence for differential errors in $X$. Therefore, we consider differential misclassification errors in both variables.

Now we would like to test whether the misclassification for $Y$ is independent of that for $X$. That is, $H_0 : D_{11} = D_{10} = D_{01} = D_{00} = 0$ versus $H_1 : \text{not } H_0$. Again we use a likelihood ratio test. The observed test statistic is 628.4, that leads to an extremely small $p$-value $< 1e^{-100}$. Therefore, we reject the null hypothesis. Taking into account the conclusion for testing differential versus nondifferential misclassification, we conclude that the misclassification errors in both variables are differential and dependent. This is consistent with the conclusion in Tang et al. (2013), where AIC was used for model comparison. It is worth noting that our model selection is based on formal test procedures, which are more objective and thus accessible for applied practitioners.

A cautionary note is advised on the potential for worse consequences of using imperfect gold standards for collecting interval validation data in some real applications. If the accuracy of the imperfect gold standards is known, one can still use our method for handling nondifferential misclassification errors in both variables. Otherwise one will have to resort to different data structures, such as replicates of $(Y^*, X^*)$, to ensure model identification.
6. Discussion and Conclusion

Different from existing work on simultaneous misclassification in both the response variable and the covariate, our main interest is to explore the bias caused by the wrong adjustment for simultaneous misclassification errors. In particular, an independent misclassification model is fitted to data generated from a dependent misclassification model. A numerical investigation is performed with different types of misclassifications (differential vs. nondifferential), different levels of proportion size of internal validation data, different amount of misclassification errors, and different strengths of dependence.

In summary, our key findings are the following. First, in contrast to the case with only one misclassified variable, the bias in OR can be more complicated in the context of dependent misclassification errors even when both variables are subject to nondifferential errors. Second, applied practitioners should be cautioned with assumptions of nondifferential and independent misclassification errors. When such assumptions are used in data analysis, one needs to carefully check them first to avoid invalid inference method and thus incorrect conclusions. Third, our simulation results indicate that adjusting the misclassification in a wrong way can produce even worse results than completely ignoring the misclassification issue (i.e., naïve method) when the proportion size of internal validation data is small (10%). Therefore, extra caution is needed when overall sample size and proportion of internal validation data are both small.

There are different lines of future work that can be built upon the current findings. First, one can further consider an alternative model accounting for dependent misclassification, especially for categorical variables. For example, differential and dependent misclassification models were considered in Salway et al. (2019). Their work focused on decomposition of $P(Y^*, X^* | Y, X)$, which was built upon the misclassification mechanism in the corresponding real application. Second, it will be interesting to carry out theoretical investigation on minimal proportion size of internal validation data to avoid the phenomenon that partial adjustment for misclassification errors is even worse than a naïve model. The results can help guide study designs (e.g., sample size for validation data).

Supplementary Materials

The supplementary materials contain two sections. The first section includes all the proofs of Supplementary Notes 1–5, as mentioned in Sections 2.2 and 2.3. The second section presents all the remaining figures other than Figures 1–3 for the simulation studies in 4.1 and 4.2.

Acknowledgments

We would like to thank the anonymous referees and the editor for their valuable suggestions that have greatly improved the presentation of the article. We would also like to thank Professor X. Joan Hu for the helpful discussion.
Funding

The first author’s work is supported by the Natural Sciences and Engineering Research Council of Canada. The second author’s work was supported by the trainee scholarship from the Saskatchewan Centre for Patient-Oriented Research. The last author’s work is partially supported by the National Science Foundation and National Institute of Health.

References

Barron, B. A. (1977), “The Effects of Misclassification on the Estimation of Relative Risk,” Biometrics, 33, 414–418. [353]
Brenner, H., Savitz, D. A., and Gefeller, O. (1993), “The Effects of Joint Misclassification of Exposure and Disease on Epidemiologic Measures of Association,” Journal of Clinical Epidemiology, 46, 1195–1202. [353,354]
Carroll, R. J., Ruppert, D., Crainiceanu, C. M., and Stefanski, L. A. (2006), Measurement Error in Nonlinear Models: A Modern Perspective, Boca Raton, FL: Chapman and Hall/CRC. [353]
Cramér, H. (1999), Mathematical Methods of Statistics, Princeton: Princeton University Press. [356]
Greenland, S., and Kleinbaum, D. G. (1983), “Correcting for Misclassification in Two-way Tables and Matched-pair Studies,” International Journal of Epidemiology, 12, 93–97. [353]
Gustafson, P. (2004), Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments, Boca Raton, FL: CRC Press. [353]
Kristensen, P. (1992), “Bias from Nondifferential but Dependent Misclassification of Exposure and Outcome,” Epidemiology, 3, 210–215. [356]
Salway, T., Plöderl, M., Liu, J., and Gustafson, P. (2019), “Effects of Multiple Forms of Information Bias on Estimated Prevalence of Suicide Attempts According to Sexual Orientation: An Application of a Bayesian Misclassification Correction Method to Data from a Systematic Review,” American Journal of Epidemiology, 188, 239–249. [353,361]
Smith, D. K., Warren, D. L., Vlahov, D., Schuman, P., Stein, M. D., Greenberg, B. L., Holmberg, S. D., and Group, H. I. V. E. R. S. (1997), “Design and Baseline Participant Characteristics of the Human Immunodeficiency Virus Epidemiology Research (HER) Study: A Prospective Cohort Study of Human Immunodeficiency Virus Infection in US Women,” American Journal of Epidemiology, 146, 459–469. [359]
Tang, L., Lyles, R. H., Ye, Y., Lo, Y., and King, C. C. (2013), “Extended Matrix and Inverse Matrix Methods Utilizing Internal Validation Data when Both Disease and Exposure Status are Misclassified,” Epidemiologic Methods, 2, 49–66. [353,359,360]
vAN Smeden, M., Lash, T. L., and Groenwold, R. H. (2020), “Reflection on Modern Methods: Five Myths about Measurement Error in Epidemiological Research,” International Journal of Epidemiology, 49, 338–347. [353]
Vogel, C., Brenner, H., Pfahlberg, A., and Gefeller, O. (2005), “The Effects of Joint Misclassification of Exposure and Disease on the Attributable Risk,” Statistics in Medicine, 24, 1881–1896. [353,354,355,356]
Yi, G. Y. (2017), Statistical Analysis with Measurement Error or Misclassification, New York: Springer. [353]