Research Article

Validation Study Methods for Estimating Odds Ratio in $2 \times 2 \times J$ Tables When Exposure is Misclassified

Bijan Nouri,¹ Najaf Zare,² and Seyyed Mohammad Taghi Ayatollahi¹

¹ Department of Biostatistics, Faculty of Medicine, Shiraz University of Medical Sciences, Shiraz 7134845794, Iran

² Department of Biostatistics, Infertility Research Center, Shiraz University of Medical Sciences, Shiraz 7134845794, Iran

Correspondence should be addressed to Najaf Zare; najafzare@sums.ac.ir

Received 27 September 2013; Revised 27 November 2013; Accepted 30 November 2013

Academic Editor: Edelmira Valero

Copyright © 2013 Bijan Nouri et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Background. Misclassification of exposure variables in epidemiologic studies may lead to biased estimation of parameters and loss of power in statistical inferences. In this paper, the inverse matrix method, as an efficient method of the correction of odds ratio for the misclassification of a binary exposure, was generalized to nondifferential misclassification and $2 \times 2 \times J$ tables. **Methods.** Simple estimates for predictive values when misclassification is nondifferential are presented. Using them, we estimated the corrected log odds ratio and its variance for $2 \times 2 \times J$ tables, using the inverse matrix method. A two-step weighted likelihood method was also developed. Moreover, we compared the matrix and inverse matrix methods to the maximum likelihood (MLE) method using a simulation study. **Results.** In all situations, the inverse matrix method proved to be more efficient than the matrix method. Matrix and inverse matrix methods for nondifferential situations are more efficient than differential misclassification. **Conclusions.** Although MLE is optimal among all of the methods, it is computationally difficult and requires programming. On the other hand, the inverse matrix method with a simple closed-form presents acceptable efficiency.

1. Introduction

In epidemiology studies, where the assessment of the relationship between exposure and outcome variables is the main goal, misclassification of exposure variable leads to biased estimate of odds ratio. In multicenter clinical trials with an increasing number of centers, the possibility of misclassification of the exposure variable and the biases induced by it will arise. Methods to correct a possibly misclassified exposure so that the strength of the association between exposure and outcome variables can be precisely assessed have been a focus of statistical and epidemiological research for over 30 years. Beginning with classic papers, the issue of misclassification on tabular data has long been recognized and their adjustment has been considered [1–5]. In 1977, the matrix method was presented by Barron in order to correct nondifferential misclassification in 2×2 tables [6]. Greenland and Kleinbaum generalized it to differential misclassification and match paired data [7]. In addition, Greenland proposed a variance estimate for the matrix method under the assumptions of differential and nondifferential misclassifications [8].

Selen, in a simulation study, showed that both the matrix and the maximum likelihood methods performed equally well [9]. In 1990, Marshall presented a more direct inverse matrix approach for correcting differential misclassifications [10]. Morrissey and Spiegelman compared the matrix and inverse matrix methods to the maximum-likelihood estimator using a grid search for 2×2 tables [11]. He found that, under the assumption of differential misclassification, the inverse matrix method was always more efficient than the matrix method. Examples of validation studies in the misclassification context in linear models are prevalent in the statistical and epidemiologic literature [12–15]. More recent works are included in the reference section [16–20].

The main impediment to the use of the inverse matrix method, despite its superiority compared to the matrix method, is that it is restricted to differential misclassification and 2×2 tables, while, in practice, it is very likely that misclassification is nondifferential or that we have a $2 \times 2 \times J$ table, in which J determines the number of centers in a multicenter clinical trial or is the levels of a confounder [21]. For example,

TABLE 1: Notation for j th stratum of a $2 \times 2 \times J$ misclassified table.

Outcome (y)	Error-prone exposure (X^*)		
	1	0	
1	$A_{1j}^* (n_{11j})$	$B_{1j}^* (n_{01j})$	n_{1j}
0	$A_{0j}^* (n_{10j})$	$B_{0j}^* (n_{00j})$	n_{0j}

when misclassification is due to recall bias, it is expected that the rates of misclassification in case and control groups are the same, leading to nondifferential misclassification. Here, we aim to further extend the focus within the inverse matrix method when misclassification is nondifferential and data is stratified on J strata.

Section 2.1 provides definition and notation for a $2 \times 2 \times J$ table with a binary error-prone exposure. For the first time, we propose estimates for negative and positive predictive values as misclassification parameters under the assumption of nondifferential misclassification, which can be used to generalize an inverse matrix method to nondifferential cases (Section 2.2). Then we will generalize an inverse matrix approach to a nondifferential assumption and situations where the data is stratified on J strata (Section 3.1). Intuitive closed-form formulae for misclassification-adjusted effect and its variance are presented. The matrix method and the likelihood method will be briefly reviewed later. In addition, we will present a new two-step method in Section 3.4, which uses corrected cell count by the matrix method to construct a weighted maximum likelihood method. Finally, the study will continue with a simulation study that compares the mean square error (MSE) of each of the presented methods to the MLE under the assumptions of differential and nondifferential misclassifications. The main text focuses on simple formulae for conditions involving a binary misclassified exposure in a case-control study, with stratification on confounders. Several assumptions were made in this work. First, the disease status and level of strata were measured accurately. Second, the methods required an error-free criterion for exposure to validate the misclassified exposure in the validation study. Finally, we assumed that the main and validation studies are independent. The formulae are illustrated with data from an often-cited case-control study of sudden infant death syndrome (SIDS) and maternal use of antibiotics during pregnancy [22].

2. Method and Material

2.1. Definitions and Notations. To begin, consider a case-control study sample; binary exposure and outcome variables were measured by two error-prone and correct methods, respectively, as well as a J level variable S (S can be a confounder or a combination of some confounders). So, the data could be classified in a $2 \times 2 \times J$ table, for which its j th stratum contains n_{1j} cases and n_{0j} controls. Because the exposure status X^* is an error-prone variable, the cross-table will be misclassified. The j th stratum of a $2 \times 2 \times J$ misclassified table and the notations that are employed in this paper are displayed in Table 1.

TABLE 2: Data layout for the validation study.

X^*	Validation study			
	$Y = 1$		$Y = 0$	
	$X = 1$	$X = 0$	$X = 1$	$X = 0$
1	$m_{111} = 29$	$m_{101} = 22$	$m_{011} = 21$	$m_{001} = 12$
0	$m_{110} = 17$	$m_{100} = 143$	$m_{010} = 16$	$m_{000} = 168$

m_{kli} .

X is supposed to be an error-free variable that shows the actual status of the exposure. The sensitivity and specificity of the error-prone diagnosis of exposure are known as misclassification parameters or misclassification rates and defined by $SE_{kj} = P(X^* = 1 | X = 1, Y = k, S = j)$ and $SP_{kj} = P(X^* = 0 | X = 0, Y = k, S = j)$ for $k = 0, 1$ and $j = 1, \dots, J$, respectively. Also one can offer misclassification rates by another set of parameters, known as positive and negative predictive values, which can be defined as $PPV_{kj} = P(X = 1 | X^* = 1, Y = k, S = j)$ and $NPV_{kj} = P(X = 0 | X^* = 0, Y = k, S = j)$ for $k = 0, 1$ and $j = 1, \dots, J$, respectively. Among different approaches developed to deal with the misclassification problem, the inverse matrix method (direct method) uses positive and negative predictive values as misclassification parameters; on the other hand, the matrix method (indirect method) and the likelihood method both apply sensitivity and specificity values instead. Let $e_{kj}^* = P(X^* = 1 | Y = k, S = j) = A_{kj}^*/n_{kj}$ represents the prevalence of error-prone exposure for $k = 0, 1$ and $j = 1, \dots, J$, and also $f_{kj}^* = 1 - e_{kj}^*$.

Both of the sets of misclassifications are unknown, and it is required to estimate them through a validation study. Thus, in addition to the sample described above known as the main study, another random sample of size m is drawn separately. This sample makes up the external validation study. In order to estimate the misclassification parameters, as well as the error-prone status of exposure X^* , the correct exposure status should be measured precisely for each subject of the validation sample by a dichotomous error-free variable X .

2.2. Estimation of Misclassification Parameters. Misclassification parameters are allowed to vary by levels of J in this approach and separate misclassification parameters for each stratum can be estimated, but if we notice that the misclassification rates are similar in strata, regardless of stratification, validation samples of strata should be combined. Due to the simplicity of notation, we will describe the case in which misclassification rates are the same across strata; otherwise, estimation of the parameters is similar. Table 2 displays the available data m_{kli} for the validation study, where subscripts $k = 0, 1$, $l = 0, 1$, and $i = 0, 1$ indicate the outcome, error-free, and error-prone exposure status, respectively. Misclassification parameters and their variances are estimated in the validation study by the following formulae:

$$SE_k = \frac{m_{k11}}{m_{k1.}} \quad SP_k = \frac{m_{k00}}{m_{k0.}},$$

$$Var(SE_k) = \frac{SE_k(1 - SE_k)}{m_{k1.}},$$

$$\begin{aligned}
\text{Var}(\text{SP}_k) &= \frac{\text{SP}_k(1 - \text{SP}_k)}{m_{k0.}}, \\
\text{PPV}_k &= \frac{m_{k11}}{m_{k.1}}, \quad \text{NPV}_k = \frac{m_{k00}}{m_{k.0}}, \\
\text{Var}(\text{NPV}_k) &= \frac{\text{NPV}_k(1 - \text{NPV}_k)}{m_{k.0}}, \\
\text{Var}(\text{NPV}_k) &= \frac{\text{NPV}_k(1 - \text{NPV}_k)}{m_{k.0}}.
\end{aligned} \tag{1}$$

In this case, a point, “,” in the subscripts stands for summation over that subscript. Misclassification is nondifferential when the misclassification rates are independent of the outcome status. In other words, when $P(X^* = 1 \mid X = 1, Y = k) = P(X^* = 1 \mid X = 1)$ and $P(X^* = 0 \mid X = 0, Y = k) = P(X^* = 0 \mid X = 0)$; otherwise, misclassification is differential. We can easily adapt estimates of sensitivity and specificity to the following nondifferential assumption:

$$\begin{aligned}
\text{SE} &= \frac{m_{.11}}{m_{.1}}, \quad \text{SP} = \frac{m_{.00}}{m_{.0}}, \\
\text{Var}(\text{SE}) &= \frac{\text{SE}(1 - \text{SE})}{m_{.1}}, \\
\text{Var}(\text{SP}) &= \frac{\text{SP}(1 - \text{SP})}{m_{.0}}.
\end{aligned} \tag{2}$$

But once we assume that misclassification is nondifferential, the sensitivity and specificity are equal given the disease status. In this case, nondifferential sensitivity and specificity do not mean that PPV and NPV are equal across strata of the outcome because PPV and NPV are functions of exposure prevalence and sensitivity and specificity. Due to this problem, the methods that apply predictive values as misclassification parameters (such as the inverse matrix method) will be restricted to the differential assumption. In order to overcome this difficulty, Morrissey and Spiegelman recommended the use of the same predictive values as those estimated by the differential misclassification assumption, which is not a reasonable approach [11]. Instead, in the case of nondifferential misclassification, we present the following estimates for predictive values:

$$\begin{aligned}
\text{PPV}_k &= \frac{(\text{SE}m_{k1.})}{m'_{k.}}, \quad \text{NPV}_k = \frac{(\text{SP}m_{k0.})}{(m_{k..} - m'_{k.})}, \\
\text{Var}(\text{PPV}_k) &= \frac{\text{PPV}_k(1 - \text{PPV}_k)}{m'_{k.}}, \\
\text{Var}(\text{NPV}_k) &= \frac{\text{NPV}_k(1 - \text{NPV}_k)}{(m_{k..} - m'_{k.})},
\end{aligned} \tag{3}$$

where $m'_{k.} = (\text{SE}m_{k1.} + (1 - \text{SP})m_{k0.})$. If sensitivity and specificity are constant and known, using the probability roles, we can estimate predictive values as follows:

$$\text{PPV}_k = \frac{\text{SE}_k(\text{SP}_k - f_k^*)}{e_k^* D_k}, \quad \text{NPV}_k = \frac{\text{SP}_k(\text{SE}_k - e_k^*)}{f_k^* D_k}. \tag{4}$$

TABLE 3: Uncorrected data from SIDS study stratified on the sex of infant.

Interview response (X^*)	Males		Females	
	$Y = 1$	$Y = 0$	$Y = 1$	$Y = 0$
Use	80	42	42	59
No use	257	261	185	218
Total	337	303	227	277

TABLE 4: Estimates of misclassification parameters and their variance.

	Differential	Nondifferential
PPV ₁	0.5686 (0.0048096)	0.6300 (0.0053002)
PPV ₀	0.6363 (0.0070127)	0.5567 (0.0061395)
NPV ₁	0.8937 (0.0005937)	0.8904 (0.0005842)
NPV ₀	0.9130 (0.0004316)	0.9168 (0.0004310)
SE ₁	0.6304 (0.0050651)	0.6024 (0.0028857)
SE ₀	0.5676 (0.0066332)	
SP ₁	0.8667 (0.0007003)	0.9014 (0.0025761)
SP ₀	0.9333 (0.0003457)	

These estimates are based on the assumption that misclassification rates are the same on male and female strata.

When equality of the misclassification parameters in J strata cannot be assumed, using distinct validation sample for each stratum and the above relations, misclassification parameters can be estimated separately.

Example 1. Table 3 exhibits the main study data from an often-cited case-control study on sudden infant death syndrome (SIDS), which has examined the relationship between maternal use of antibiotics during pregnancy and the odds of SIDS [22]. The table is classified on two strata ($J = 2$) according to the sex of infants, a common risk factor for SIDS. Drug use error-prone measurement was an interview response X^* , and it was validated by medical records X . A separate sample from a validation study with dual exposure measurement (both error-prone and error-free) is presented in Table 2. Using (1)–(3) and the validation data, separate estimates with regard to both differential and nondifferential misclassification assumptions for sensitivity, specificity, and predictive values are presented in Table 4. Two cells in the last column of Table 4 are blank because, for nondifferential misclassification, sensitivity and specificity are the same for case and control groups.

3. Estimate of the Actual Association

3.1. Inverse Matrix Method. As mentioned in the introduction, the inverse matrix method was proposed by Marshall to correct misclassified 2×2 tables [10]. He restricted the use of his method to the assumption of differential misclassification. According to the estimates we present for predictive values in (3), the inverse matrix method can be developed to the differential assumption. Now, we generalize this approach to $2 \times 2 \times J$ tables with misclassified exposure and either differential or nondifferential assumptions. Estimates for

correct proportions of exposed and unexposed subjects in the j th stratum of the main study are $e_{kj} = \text{PPV}_{kj}e_{kj}^* + (1 - \text{NPV}_{kj})f_{kj}^*$ and $f_{kj} = \text{NPV}_{kj}f_{kj}^* + (1 - \text{PPV}_{kj})e_{kj}^*$, respectively. These equations can be written in the matrix form as

$$\begin{bmatrix} e_{1j} \\ f_{1j} \\ e_{0j} \\ f_{0j} \end{bmatrix} = \begin{bmatrix} \text{PPV}_{1j} & 1 - \text{NPV}_{1j} & 0 & 0 \\ 1 - \text{PPV}_{1j} & \text{NPV}_{1j} & 0 & 0 \\ 0 & 0 & \text{PPV}_{0j} & 1 - \text{NPV}_{0j} \\ 0 & 0 & 1 - \text{PPV}_{0j} & \text{NPV}_{0j} \end{bmatrix} \times \begin{bmatrix} e_{1j}^* \\ f_{1j}^* \\ e_{0j}^* \\ f_{0j}^* \end{bmatrix}. \quad (5)$$

The biased log odds ratio estimate of the j th stratum is $\theta_j^* = \log(A_{1j}^*B_{0j}^*/A_{0j}^*B_{1j}^*)$, with asymptotic variance estimate $V_j^* = \sum_k (1/A_{kj}^* + 1/B_{kj}^*)$ where $k = 0, 1$. Corrected values in (5) are used to build the inverse matrix corrected log-odds-ratio $\theta_j = \log(e_{1j}f_{0j}/e_{0j}f_{1j})$ where $e_{kj} = P(X = 1 | Y = k, S = j)$ representing the estimate of actual prevalence of exposure for $k = 0, 1$ and $j = 1, \dots, J$, and also $f_{kj} = 1 - e_{kj}$. Considering binomial distribution for A_{kj}^* , we derived the asymptotic variance for θ_j using the delta method as

$$V_{jj} = \text{Var}(\theta_j) = \sum_{k=0,1} \frac{(d_{kj}^2 V_{e_{kj}^*} + e_{kj}^{*2} V_{\text{PPV}_{kj}} + f_{kj}^{*2} V_{\text{NPV}_{kj}})}{(f_{kj}^2 e_{kj}^2)}, \quad (6)$$

where the letter V stands for an abbreviation for variance and $d_{kj}^2 = \text{PPV}_{kj} + \text{NPV}_{kj} - 1$. If nondifferential misclassification is assumed, instead of the estimated positive and negative predictive values and their variance estimate in (1), corresponding values in (3) must be considered substitutes in expression (6) and also the corrected proportions in the disease group are no longer independent of the corrected proportions in the control group. Consequently, twice the covariance between log odds (7) must be subtracted from expression (6). Consider

$$\frac{V_{\text{SE}_j}}{f_{0j}f_{1j}D_j^2} + \frac{V_{\text{SP}_j}}{e_{0j}e_{1j}D_j^2}. \quad (7)$$

If in the validation study for each stratum a separate misclassification-rate is estimated, the log odds ratios are independent; otherwise, using the same misclassification parameters for all the strata leads to covariance (8) between θ_j and θ_i . Consider

$$V_{ij} = \sum_{k=0,1} \frac{(V_{\text{SE}_k}/f_{kj}f_{ki} + V_{\text{SP}_k}/e_{kj}e_{ki})}{D_k^2}. \quad (8)$$

If the same misclassification parameters for all the strata are estimated by assuming nondifferential misclassification, the covariance estimate for θ_j and θ_i is given by

$$V_{ij} = \sum_{k=0,1} \sum_{h=0,1} (-1)^{h+k} \frac{(V_{\text{SE}}/f_{kj}f_{hi} + V_{\text{SP}}/e_{kj}e_{hi})}{D^2}. \quad (9)$$

Let V be the symmetric J by J variance-covariance matrix of the J estimated odds-ratio, with ij th element V_{ij} . Then $\theta_w = (\theta'V^{-1}\mathbf{1})V_w$ is the minimum variance weighted average of the estimated log-odds ratios, where $V_w = 1/(\mathbf{1}'V^{-1}\mathbf{1})$, and also $\mathbf{1}$ and θ are J -vector of ones and estimated odds ratios, respectively. For large strata sample, θ_w is a consistent uniformly asymptotic normal estimator for the actual log odds ratio, with consistent variance estimator V_w . Hence, construction of a corrected Wald test of no association will be possible using the z statistic $\theta_w/\sqrt{V_w}$, with a corresponding corrected $1 - \alpha$ percent confidence interval for the actual log odds ratio as $\theta_w \pm z_{t=1-\alpha/2}\sqrt{V_w}$, where z_t is the $t = 1 - \alpha/2$ percentile of a standard normal distribution.

In order to evaluate the homogeneity of θ_j , we can utilize an approximate $J - 1$ degree of freedom Chi-square statistics $X^2 = \theta'V^{-1}\theta - V_w^{-1}\theta_w^2$. Large values of this statistic indicate loss of homogeneity and so θ_w will be an inadequate summary of the association.

3.2. Matrix Method. Now we briefly outline the matrix method. In this method sensitivity and specificity are used as misclassification parameters to estimate the log-odds ratio. The matrix method uses the equations $A_{kj}^* = \text{SE}_{kj}A_{kj} + (1 - \text{SP}_{kj})B_{kj}$ and $B_{kj}^* = (1 - \text{SE}_{kj})A_{kj} + \text{SP}_{kj}B_{kj}$ to estimate correct cell counts in the j th stratum; in matrix form these equations can be written as

$$\begin{bmatrix} A_{1j} \\ B_{1j} \\ A_{0j} \\ B_{0j} \end{bmatrix} = \begin{bmatrix} \text{SE}_{1j} & 1 - \text{SP}_{1j} & 0 & 0 \\ 1 - \text{SE}_{1j} & \text{SP}_{1j} & 0 & 0 \\ 0 & 0 & \text{SE}_{0j} & 1 - \text{SP}_{0j} \\ 0 & 0 & 1 - \text{SE}_{0j} & \text{SP}_{0j} \end{bmatrix}^{-1} \times \begin{bmatrix} A_{1j}^* \\ B_{1j}^* \\ A_{0j}^* \\ B_{0j}^* \end{bmatrix}. \quad (10)$$

The matrix corrected log-odds-ratio is $\theta_j = \log(e_{1j}f_{0j}/e_{0j}f_{1j})$, $e_{kj} = A_{kj}/n_{kj}$ and $f_{kj} = 1 - e_{kj}$. Considering binomial

distribution for A_{kj}^* , Greenland provided the asymptotic variance for θ_j , given by

$$V_{jj} = \text{Var}(\theta_j) = \sum_{k=0,1} \frac{(V_{SE_{kj}}/f_{kj}^2 + V_{SP_{kj}}/e_{kj}^2 + e_{kj}^* f_{kj}^*/n_{kj} e_{kj}^2 f_{kj}^2)}{(D_{kj}^2)}, \quad (11)$$

where $D_{kj} = SE_{kj} + SP_{kj} - 1$. The other components of the variance-covariance matrix are similar to what we derived for the inverse matrix method.

This method has two limitations: when the sum of sensitivity and specificity is equal to one, the corrected count estimates are undefined because the misclassification matrix is singular in (10); if the sum is less than one, negative estimates for corrected cell counts occur.

3.2.1. Direct Likelihood Method. Some authors use the maximum likelihood estimate (MLE) to estimate the actual odds ratio and test the exposure-disease association. Although this method can be the most efficient method of correction, it does not have a close form and requires an iterative solution to a nonlinear set of equations under some constraints. If the error-free exposure variable X was available, we would be able to estimate the covariate-adjusted odds ratio by the following multiple logistic regression model:

$$\text{logit} [\Pr(Y = 1 \mid X, C)] = \alpha + \theta x + \sum_{j=1}^{J-1} \gamma_j c_j. \quad (12)$$

However, we just have access to the error-prone binary variable X^* . We assume that the covariates $C_j (j = 1, \dots, J-1)$ are dummy variables that determine to which stratum a subject belongs. If the subject belongs to the j th stratum $C_j = 1$, as mentioned before, θ indicates overall log odds ratio. To obtain the ML estimate for unknown parameter θ , we should numerically maximize the following log likelihood function with respect to θ . The approximate standard error for this estimate is available by inverting the observed information matrix. Consider

$$l(\theta) = \sum_j \sum_k \sum_i n_{ikj} \ln(\Pr(X^* = i, Y = k \mid C_j = c)), \quad (13)$$

where n_{ikj} was presented in Table 1; note that $\Pr(X^*, Y \mid C = c)$ in expression (13) is the observed data likelihood contribution in the likelihood function which can be obtained as follows:

$$\sum_{x=0}^1 \Pr(X^* = i \mid X = l, Y = k, C = c) \times \Pr(Y = k \mid X = l, C = c) \Pr(X = l \mid C = c), \quad (14)$$

where the first term in (14) can be obtained using the estimated sensitivity and specificity in (1) or (2) according to whether misclassification is differential or nondifferential, the second term displays the logistic model (12), and finally the last term is a nuisance parameter that can be modeled via a second logistic regression model of X on C .

3.3. A Weighted Likelihood Method. Now, we briefly present a two-step method which is a combination of the inverse matrix method and the likelihood method. In the first step, we use the inverse matrix method to correct misclassified tables, and in the second step, we utilize corrected cell count in the first step that is assigned to (x, y, j) log-likelihood contribution as weights. Formally, the weighted log-likelihood is given by

$$l(\theta) = \sum_{j=1}^J \sum_{k=0}^1 \sum_{l=0}^1 N_{lkj} l_{lkj}(\theta), \quad (15)$$

where the N_{xyk} represents the corrected data cell count by the inverse matrix method and subscripts $l = 0, 1, k = 0, 1$, and $j = 1, \dots, J$ indicate error-free exposure, outcome, and strata number, respectively. N_{lkj} represents the corrected data cell count by the inverse matrix method and $l_{lkj}(\theta)$ is its log-likelihood contribution. This weighted method is simpler than the previous likelihood method and directly uses (x, y, j) log-likelihood contribution instead of (x^*, y, j) . The variance of corrected log-OR can be obtained by inverting the Fisher information matrix.

3.4. Combining the Results. Now, we need an approach to combine the results from the main study and the internal validation study. We emphasize that, for subjects in the validation study, a conventional analysis using an error-free exposure can be done to obtain a log-OR estimate θ_v with variance estimate V_v . The estimate of the effect for the main study is biased and must be adjusted using the earlier formulae in order to obtain a corrected log-OR estimate θ_m with variance estimate V_m . A weighted method to combine the log-OR from the two samples can be constructed as follows:

$$\theta_c = \frac{(\theta_m V_v + \theta_v V_m)}{(V_m + V_v)}. \quad (16)$$

θ_c has variance estimate $V_c = V_m V_v / (V_m + V_v)$. The weighted method proceeds well if θ_v and θ_m estimate a common value, and this is achieved when samples are randomly selected from the population, and the misclassification parameter estimates are unbiased.

Example 2. For the data in Table 3, we have misclassified OR = 1.124 and $\ln(\text{OR}) = 0.1174(0.02357)$, in which the term in the bracket represents the estimated variance. Suppose that the investigator assumes that misclassification rates are equal in both the strata of male and female infants. Under the nondifferential assumption and using the parameters estimated in Table 4, the matrix method yields $\widehat{\text{OR}} = 1.26$ and $\ln(\widehat{\text{OR}}) = 0.2311(0.04305)$, and the inverse matrix method yields $\widehat{\text{OR}} = 1.237$ and $\ln(\widehat{\text{OR}}) = 0.213(0.04103)$. In contrast, we supposedly have differential estimates for misclassification rates in Table 2; thus, the results from the matrix and inverse matrix method will be $\ln(\widehat{\text{OR}}) = 0.2678(0.05147)$ and $\ln(\widehat{\text{OR}}) = 0.2471(0.04012)$, respectively.

Applying the direct ML (13) to the data in Table 3 yields $\hat{\beta} = \ln(\widehat{OR}) = 0.00585(0.02418)$ and $\hat{\beta} = \ln(\widehat{OR}) = 0.505(0.0312)$ under the differential and nondifferential assumptions, respectively. Note the dramatic shift in the implication of assuming a differential misclassification. The application of the two-step weighted likelihood approach produces $\hat{\beta} = \ln(\widehat{OR}) = 0.3027(0.048309)$ and $\hat{\beta} = \ln(\widehat{OR}) = 0.3047(0.04831)$ given the differential and nondifferential cases, respectively.

4. Simulation Studies

We conducted a simulation study to compare corrected log-odds-ratios obtained from both matrix and inverse matrix methods and also to assess the appropriateness of the maximum likelihood and weighted maximum likelihood for $2 \times 2 \times J$ tables. In order to mimic the SIDS example, a total of 1000 sets of data were generated with a binary outcome (Y) and also two other binary variables as error-free exposure status and stratification indicator (X and S) and a total sample size of 1144. Correct cell counts for each table were generated with regard to the following simulation conditions: prevalence of $P(S = 1) = 0.5$ and conditional prevalence of exposure are $P(X = 1 | S = 1) = 0.5$ and $P(X = 1 | S = 0) = 0.5$. In order to generate the response variable, the binary distribution with parameter $\pi_{ij} = \exp(\alpha + \theta x_i + \gamma S_j) / (1 + \exp(\alpha + \theta x_i + \gamma S_j))$ was used, where we set the effect parameters for all the simulation scenarios as $\alpha = -1$, $\theta = 1.71$ and $\gamma = 1$. The misclassified exposure (X^*) was generated assuming numerous nondifferential conditions. In order to estimate the misclassification parameters for every one of the simulated data sets, three scenarios of 20%, 30%, and 40% of each of their strata were taken as internal validation samples separately. The programming code could be provided via request from the authors.

The last four columns of Table 5 represent different scenarios related to cases in which nondifferential misclassifications vary across strata, while the four columns before them display various situations in which nondifferential misclassifications are constant across strata; analysis of each column is in accordance with the manner in which that column was generated. For the purpose of comparing the different methods, the mean square error (MSE) measure was used in order to take into account the bias and variance of estimators simultaneously. As you can see, the MLE estimator is more efficient than other estimator methods in all scenarios. Our simulation study showed that for nondifferential misclassification the performance of these inverse matrix methods and likelihood method is very close but is not equivalent. The inverse matrix method is significantly more precise than the matrix method. Performance of the likelihood and weighted likelihood methods is relatively the same under both assumptions. The efficiency of all estimators would be improved by increasing the validation study sample size, but this improvement would be more prominent in the weighted likelihood method, as the efficiency of the estimator will be approximately doubled by increasing the validation study sample size.

5. Discussion and Conclusions

The misclassification of exposure is an issue of broad concern in epidemiological research studies, and considerable amount of pieces of literature discuss adjusting inferences on exposure-disease relationships regarding such misclassification. Matrix method, inverse matrix method, likelihood method, Bayesian method, and SIMEX method are some of the general tools for performing such adjustments. Our main focus was on the development of the inverse matrix method to nondifferential misclassification and $2 \times 2 \times J$ tables. At least in the context of misclassified binary exposure, this paper has illustrated several positive attributes of the inverse matrix method. First, to assess the diagnostic tests, the importance of predictive values (rather than sensitivity and specificity) is well known. The inverse matrix method, as a direct method, uses the predictive values. We presented simple closed-form estimates for positive and negative predictive values when misclassification is nondifferential. Second, the simulation study showed that the inverse matrix method estimator is practically more precise than the matrix method in all scenarios.

We also examined the effect of the parameters controlled by the researcher on the MSEs of these approaches. We found that the optimal estimator depends on the size of the validation study. This was due to the fact that, in some cases where the validation sample is large, a good part of the information about θ can come from the validation study. Morrissey and Spiegelman noted this phenomenon previously from the comparison of these methods for 2×2 tables. These researchers showed that the efficiency of estimators depends more on the relative size of the validation study than on the case-control ratio [11]. In another study, Greenland noted that the samples from which the misclassification parameters are estimated must be large enough to assure the approximate normality of SE, SP, PPV, and NPV [8]. Lyles noted clearly that under the differential misclassification setting, the inverse matrix method is equivalent to the maximum likelihood approach [19]. There are closed-form solutions for both, which are exactly the same. For this reason, both approaches should be equally more efficient than the matrix method.

The methods described in this study handle the misclassification of exposure, but they are easily adapted for disease misclassification. The study depends on some assumptions stated at the end of the introduction. Further studies seem to be necessary to compare and extend these methods, as those assumptions are not met. This would make sense when the outcome status and exposure status are simultaneously misclassified and when there is not a gold standard to validate the exposure.

In conclusion, when trying to make a decision about which method to be used, it can be helpful to consider the following concept. Although the MLE method has minimum MSE in all the scenarios, it can be computationally difficult and does not have a simple closed-form expression like the matrix and inverse matrix methods. When misclassification is non-differential, the inverse matrix method performs very well. Even when the size of the validation study is large enough, it can perform equal to the likelihood method.

TABLE 5: Simulation results for a $2 \times 2 \times 2$ table when binary exposure is misclassified.

Validation sample rate	Method	Strata 1												Strata 2											
		Se ₁₁ = 0.90 Sp ₁₁ = 0.90 Se ₁₀ = 0.90 Sp ₁₀ = 0.90	Se ₁₁ = 0.85 Sp ₁₁ = 0.90 Se ₁₀ = 0.85 Sp ₁₀ = 0.90	Se ₁₁ = 0.80 Sp ₁₁ = 0.85 Se ₁₀ = 0.80 Sp ₁₀ = 0.85	Se ₁₁ = 0.75 Sp ₁₁ = 0.80 Se ₁₀ = 0.75 Sp ₁₀ = 0.80	Se ₁₁ = 0.90 Sp ₁₁ = 0.90 Se ₁₀ = 0.90 Sp ₁₀ = 0.90	Se ₁₁ = 0.85 Sp ₁₁ = 0.90 Se ₁₀ = 0.85 Sp ₁₀ = 0.90	Se ₁₁ = 0.80 Sp ₁₁ = 0.85 Se ₁₀ = 0.80 Sp ₁₀ = 0.85	Se ₂₁ = 0.90 Sp ₂₁ = 0.90 Se ₂₀ = 0.90 Sp ₂₀ = 0.90	Se ₂₁ = 0.85 Sp ₂₁ = 0.85 Se ₂₀ = 0.85 Sp ₂₀ = 0.85	Se ₂₁ = 0.80 Sp ₂₁ = 0.80 Se ₂₀ = 0.80 Sp ₂₀ = 0.80	Se ₂₁ = 0.75 Sp ₂₁ = 0.80 Se ₂₀ = 0.75 Sp ₂₀ = 0.80	Se ₂₁ = 0.85 Sp ₂₁ = 0.85 Se ₂₀ = 0.85 Sp ₂₀ = 0.85	Se ₂₁ = 0.80 Sp ₂₁ = 0.80 Se ₂₀ = 0.80 Sp ₂₀ = 0.80	Se ₂₁ = 0.70 Sp ₂₁ = 0.80 Se ₂₀ = 0.70 Sp ₂₀ = 0.80	Se ₂₁ = 0.90 Sp ₂₁ = 0.95 Se ₂₀ = 0.90 Sp ₂₀ = 0.95									
n _v = 0.2	Matrix	1.717 (0.0298)	1.711 (0.0325)	1.671 (0.0400)	1.652 (0.0483)	1.720 (0.0301)	1.713 (0.0338)	1.682 (0.0418)	1.715 (0.0375)																
	Inverse matrix	1.724 (0.0232)	1.723 (0.0234)	1.732 (0.0238)	1.686 (0.0240)	1.727 (0.0218)	1.731 (0.0232)	1.634 (0.0252)	1.762 (0.0233)																
	Likelihood	1.723 (0.0146)	1.725 (0.0155)	1.730 (0.0176)	1.689 (0.0182)	1.724 (0.0165)	1.730 (0.0182)	1.682 (0.0203)	1.724 (0.0187)																
	Weighted likelihood	1.737 (0.0560)	1.726 (0.0554)	1.730 (0.0561)	1.685 (0.0539)	1.738 (0.0534)	1.730 (0.0532)	1.691 (0.0560)	1.721 (0.0544)																
		1.718 (0.0247)	1.720 (0.0263)	1.721 (0.0300)	1.659 (0.0358)	1.721 (0.0272)	1.711 (0.0330)	1.694 (0.0362)	1.718 (0.0335)																
n _v = 0.3	Matrix	1.720 (0.0188)	1.723 (0.0194)	1.723 (0.0237)	1.678 (0.0310)	1.723 (0.0223)	1.729 (0.0240)	1.673 (0.0322)	1.728 (0.0232)																
	Inverse matrix	1.719 (0.0140)	1.719 (0.0144)	1.716 (0.0178)	1.671 (0.0174)	1.711 (0.0145)	1.722 (0.0181)	1.696 (0.0181)	1.723 (0.0173)																
	Likelihood	1.727 (0.0372)	1.740 (0.0383)	1.745 (0.0431)	1.682 (0.0364)	1.721 (0.0375)	1.728 (0.0388)	1.624 (0.0443)	1.733 (0.0362)																
	Weighted likelihood	1.714 (0.0214)	1.719 (0.0228)	1.713 (0.0249)	1.662 (0.0293)	1.718 (0.0224)	1.712 (0.0235)	1.697 (0.0250)	1.711 (0.0233)																
		1.712 (0.0162)	1.725 (0.0167)	1.723 (0.0140)	1.684 (0.0178)	1.715 (0.0166)	1.727 (0.0165)	1.692 (0.0162)	1.729 (0.0165)																
n _v = 0.4	Matrix	1.710 (0.0132)	1.722 (0.0140)	1.716 (0.0147)	1.673 (0.0163)	1.710 (0.0131)	1.716 (0.0140)	1.688 (0.0142)	1.711 (0.0162)																
	Inverse matrix	1.726 (0.0280)	1.724 (0.0278)	1.727 (0.0275)	1.678 (0.0277)	1.729 (0.0281)	1.728 (0.0282)	1.674 (0.0275)	1.729 (0.0273)																
	Likelihood																								
	Weighted likelihood																								

Numbers in each cell reflect mean of adjusted odds ratios (MSE) based on 1000 simulated data sets, with true OR = 1.71.

The two-step likelihood method has a simpler form than the likelihood method and can perform for nondifferential misclassification like the matrix and inverse matrix methods.

Conflict of Interests

The authors declare that they have no conflict of interests.

Authors' Contribution

Bijan Nouri and Najaf Zare were responsible for the design, simulation, and interpretation. Seyyed Mohammad Taghi Ayatollahi supervised the study and interpreted the results. All the authors read and approved the final paper.

Acknowledgments

This work was a part of the Ph.D. thesis of Bijan Nouri and was supported by Shiraz University of Medical Sciences, Shiraz, Iran. The authors are also thankful to the reviewers for their invaluable comments.

References

- [1] I. Bross, "Misclassification in 2×2 tables," *Biometrics*, vol. 10, no. 4, pp. 478–486, 1954.
- [2] I. D. J. Bross, "Spurious effects from an extraneous variable," *Journal of Chronic Diseases*, vol. 19, no. 6, pp. 637–647, 1966.
- [3] J. Cornfield, W. Haenszel, E. C. Hammond, A. M. Lillienfeld, M. B. Shimkin, and E. L. Wynder, "Smoking and lung cancer: recent evidence and a discussion of some questions," *Journal of the National Cancer Institute*, vol. 22, pp. 173–203, 1959.
- [4] G. G. Koch, "The effect of non-sampling errors on measures of association in 2×2 contingency tables," *Journal of the American Statistical Association*, vol. 64, no. 327, pp. 852–863, 1969.
- [5] A. Tenenbein, "A double sampling scheme for estimating from binomial data with misclassifications," *Journal of the American Statistical Association*, vol. 65, no. 331, pp. 1350–1361, 1970.
- [6] B. A. Barron, "The effects of misclassification on the estimation of relative risk," *Biometrics*, vol. 33, no. 2, pp. 414–418, 1977.
- [7] S. Greenland and D. Kleinbaum, "Correcting for misclassification in two-way tables and matched-pair studies," *International Journal of Epidemiology*, vol. 12, no. 1, pp. 93–97, 1983.
- [8] S. Greenland, "Variance estimation for epidemiologic effect estimates under misclassification," *Statistics in Medicine*, vol. 7, no. 7, pp. 745–757, 1988.
- [9] J. Selen, "Adjusting for errors in classification and measurement in the analysis of partly and purely categorical data," *Journal of the American Statistical Association*, vol. 81, no. 393, pp. 75–81, 1986.
- [10] R. J. Marshall, "Validation study methods for estimating exposure proportions and odds ratios with misclassified data," *Journal of Clinical Epidemiology*, vol. 43, no. 9, pp. 941–947, 1990.
- [11] M. J. Morrissey and D. Spiegelman, "Matrix methods for estimating odds ratios with misclassified exposure data: extensions and comparisons," *Biometrics*, vol. 55, no. 2, pp. 338–344, 1999.
- [12] O. Davidov, D. Faraggi, and B. Reiser, "Misclassification in logistic regression with discrete covariates," *Biometrical Journal*, vol. 45, no. 5, pp. 541–553, 2003.
- [13] X. Liu and K.-Y. Liang, "Adjustment for non-differential misclassification error in the generalized linear model," *Statistics in Medicine*, vol. 10, no. 8, pp. 1197–1211, 1991.
- [14] X. Luan, W. Pan, S. G. Gerberich, and B. P. Carlin, "Does it always help to adjust for misclassification of a binary outcome in logistics regression?" *Statistics in Medicine*, vol. 24, no. 14, pp. 2221–2234, 2005.
- [15] L. S. Magder and J. P. Hughes, "Logistic regression when the outcome is measured with uncertainty," *The American Journal of Epidemiology*, vol. 146, no. 2, pp. 195–203, 1997.
- [16] R. J. Carroll, D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu, *Measurement Error in Nonlinear Models: A Modern Perspective*, Chapman & Hall, New York, NY, USA, 2010.
- [17] H. Chu, Z. Wang, S. R. Cole, and S. Greenland, "Sensitivity analysis of misclassification: a graphical and a Bayesian approach," *Annals of Epidemiology*, vol. 16, no. 11, pp. 834–841, 2006.
- [18] S. Greenland, "Maximum-likelihood and closed-form estimators of epidemiologic measures under misclassification," *Journal of Statistical Planning and Inference*, vol. 138, no. 2, pp. 528–538, 2008.
- [19] R. H. Lyles, "A note on estimating crude odds ratios in case-control studies with differentially misclassified exposure," *Biometrics*, vol. 58, no. 4, pp. 1034–1037, 2002.
- [20] R. H. Lyles and J. Lin, "Sensitivity analysis for misclassification in logistic regression via likelihood methods and predictive value weighting," *Statistics in Medicine*, vol. 29, no. 22, pp. 2297–2309, 2010.
- [21] A. M. Jurek, S. Greenland, and G. Maldonado, "Brief report: how far from non-differential does exposure or disease misclassification have to be to bias measures of association away from the null?" *International Journal of Epidemiology*, vol. 37, no. 2, pp. 382–385, 2008.
- [22] J. F. Kraus, S. Greenland, and M. Bulterys, "Risk factors for sudden infant death syndrome in the US collaborative perinatal project," *International Journal of Epidemiology*, vol. 18, no. 1, pp. 113–120, 1989.