Bias-adjusted exposure odds ratio for misclassified data

T Lee

Citation

Abstract
If a dichotomous exposure variable is misclassified in a case-control study, a bias-adjusted exposure odds ratio with its asymptotic variance is presented to account for the misclassification bias. A simple, yet powerful, method is given to calculate the true sensitivity and specificity based only on the data available in the main study, regardless of whether a validation sample is available or not. Two practical examples without or with validation data are given to illustrate how to calculate first the true sensitivity and specificity for cases and controls and then the bias-adjusted exposure odds ratio with its 95% confidence interval.

ABBREVIATIONS
BAOR Bias-adjusted [exposure] odds ratio
CI Confidence interval
COR Crude [exposure] odds ratio

INTRODUCTION
In the realm of epidemiology the problem of misclassification has been thoroughly studied. In practical applications, the exposure misclassification mainly occurs when proxy respondents are used in the survey interview to classify the subject’s exposure status. For example, in a study of identifying the possible etiologic factors for Alzheimer’s disease, information were uniformly obtained only from close family members, usually spouse, because of the patient’s mental impairment. Historically, this problem was first studied in and later included other related issues that were investigated by other people. Epidemiologic examples about the effect of misclassification bias were also widely studied. See, for example, .

So far, all proposed methods for correcting the misclassification bias either require a second validation sample to estimate the sensitivity and specificity of the classified procedure or conduct a conventional/probabilistic sensitivity analysis. No methods available in the literature are able to calculate the true sensitivity and specificity. The aim of this paper is to present a method to calculate the true sensitivity and specificity from the data in the main study only, regardless of whether a validation sample is available.

BACKGROUND
Consider a case-control study in which there is no disease misclassification, but misclassification has occurred in determining the subject’s exposure status. First, three random variables, E, E' and D, are defined as follows:

\[ E = 1 \text{ if a subject is truly exposed, } 0 \text{ otherwise, } E' = 1 \text{ if a subject is classified as exposed, } 0 \text{ otherwise. } D = 1 \text{ if a subject belongs to the case group, } 0 \text{ otherwise.} \]

Note that E' is a surrogate classification variable for the exposure variable E and D is a disease variable. Let \( p \) and \( p_1 \) denote, respectively, the true proportions of subjects in the control and the case population, who are exposed to a certain risk factor under study. The probability distributions for cases and controls are given, respectively, by the first and second column of Table 1.

Table 1: The true distributions of the cell probabilities for the cases and controls

<table>
<thead>
<tr>
<th>True exposure status</th>
<th>Group of subjects</th>
<th>D = 1 (Cases)</th>
<th>D = 0 (Controls)</th>
</tr>
</thead>
<tbody>
<tr>
<td>E = 1 (exposed)</td>
<td></td>
<td>( p_1 )</td>
<td>( p_0 )</td>
</tr>
<tr>
<td>E = 0 (unexposed)</td>
<td></td>
<td>( q_1 )</td>
<td>( q_0 )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( p_1 + q_1 = 1 )</td>
<td>( p_0 + q_0 = 1 )</td>
</tr>
</tbody>
</table>

As a measure of the relative risk in case-control studies, the [exposure] odds ratio of exposed versus unexposed is given by (\( \psi \))
Bias-adjusted exposure odds ratio for misclassified data

Figure 2

\[ R = \frac{p_0 \cdot q_0}{p_1 \cdot q_1} \]  

(1)

where \( 0 < p < 1 \) (\( q = 1 - p \)) and \( 0 < p_i < 1 \) (\( q_i = 1 - p_i \)) are defined in Table 1.

Suppose that \( n \) controls and \( n_i \) cases are sampled with the positive count frequencies \( n_{ij} \), \( i, j = 0, 1 \). By the method of moments, we obtain from Table 2

Figure 3

\[ \hat{p}_0 = \frac{n_{00}}{n_0}, \quad \hat{q}_0 = 1 - \hat{p}_0, \]  

(2)

and

Figure 4

\[ \hat{p}_1 = \frac{n_{11}}{n_1}, \quad \hat{q}_1 = 1 - \hat{p}_1, \]  

(3)

where \( \hat{p}_i \) of equation 2 and \( \hat{p}_i \) of equation 3 are the traditional sample estimates for the prevalence among controls and cases, respectively, and are unbiased estimators for the true \( p \) and \( p_i \) in Table 1, provided that there is no misclassification on the exposure variable \( E \).

Table 2: The observed cell counts for a case-control study.

<table>
<thead>
<tr>
<th>Classified exposure status</th>
<th>Group of subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td>( E = 1 ) (exposed)</td>
<td>( D = 1 ) (Cases)</td>
</tr>
<tr>
<td>( E = 0 ) (unexposed)</td>
<td>( n_{10} )</td>
</tr>
<tr>
<td>Total sample size</td>
<td>( n_1 )</td>
</tr>
</tbody>
</table>

However, \( \hat{p}_i \), \( i = 0, 1 \), of equations 2-3 are no longer unbiased estimators for \( p \) of Table 1 whenever a surrogate variable \( E' \) of the exposure variable \( E \) for the study subjects is misclassified (\( \phi \)). Indeed, once the exposure misclassification has occurred, it is easily shown that

Figure 6

\[ E(\hat{p}_i) = \frac{p_i \cdot (\phi_i + \psi_i - 1) + 1 - \psi_i}{1 + \psi_i}, \quad i = 0, 1, \]  

(4)

where \( \phi \) and \( \psi \), \( i = 0, 1 \), called bias parameters (\( \phi \)) denote sensitivity and specificity probabilities for controls and cases, respectively, and are defined by (\( \phi \))

Figure 7

\[ \phi_i = \Pr(\hat{E} = 1 | E = 1, D = i), \]  

(5)

and

Figure 8

\[ \psi_i = \Pr(\hat{E} = 0 | E = 0, D = i). \]  

(6)

Moreover, if \( n_i \cdot \hat{p}_i \)'s, \( i = 0, 1 \), are assumed to follow binomial distributions with means \( n_i \cdot \{ \hat{p}_i \cdot (\phi_i + \psi_i - 1) - \hat{q}_i \} \), the variances of \( \hat{p}_i \)'s, \( i = 0, 1 \), are given by

Figure 9

\[ \text{var}(\hat{p}_i) = \frac{n_i \cdot (\phi_i \cdot (\phi_i + \psi_i - 1) + 1 - \psi_i)}{n_i \cdot \{ \hat{p}_i \cdot (\phi_i + \psi_i - 1) - \hat{q}_i \}}, \]  

(7)

From equation 4, it is easily seen that \( \hat{p}_i \)'s, \( i = 0, 1 \) are no longer unbiased estimators of \( p_i \) unless there is no exposure misclassification for both cases and controls, that is, \( \phi_i = \psi_i = 1 \), \( i = 0, 1 \). As a result, the bias unavoidably appears in the crude [exposure] odds ratio given by

Figure 10

\[ \hat{R} = \frac{\hat{p}_1 \cdot \hat{q}_0}{\hat{p}_0 \cdot \hat{q}_1} = \frac{n_{00} \cdot n_{11}}{n_{01} \cdot n_{10}}, \]  

(8)

since it does not account for the misclassification bias. This motivates epidemiologists and statisticians to search for the corrected [exposure] odds ratio which is able to account for the misclassification bias (\( \phi \)). In this paper, an estimator, called bias-adjusted [exposure] odds ratio, is proposed which is able to account for the misclassification bias in the estimation of the true \( R \) of equation 1.

METHOD

In epidemiology, an exposure misclassification is said to be non-differential if sensitivity and specificity are the same for cases and controls, that is, classification rates are independent of the disease; otherwise, the exposure misclassification is called differential. Because non-differential misclassification is a special case of differential misclassification, I only consider differential misclassification in my derivation.

By using equations 2-4 with an approximation, \( E(\hat{p}_i) = \hat{p}_i \), it is easily shown that for \( i = 0, 1 \),

Figure 11

\[ p_i^* = (\psi_i - \hat{q}_i) / \Delta_i, \]  

(9)
Bias-adjusted exposure odds ratio for misclassified data

Figure 12

\[ q_i^* = (\delta_i - \hat{p}_i)/\Delta_i, \quad i = 0, 1 \]  

(10)

are unbiased estimators, respectively, for \( p_i \) and \( q_i \), conditioned on that both \( \delta_i \) and \( \delta_i \), \( i = 0, 1 \), are known, where \( \delta_i \) is given by

Figure 13

\[ \Delta_i = \delta_i + \psi_i - 1, \quad i = 0, 1 \]  

(11)

Clearly, equation 11 must not equal to zero; otherwise, equations 9-10 are undefined.

Now, \( p_i^* \), of equation 9 (or \( q_i^* \), of equation 10) is said to be admissible if equations 9-10 are positive numbers between 0 and 1. In addition, \( \delta_i \) and \( \delta_i \), \( i = 0, 1 \), are said to be feasible if \( \delta_i \), \( \delta_i \), \( i = 0, 1 \), and \( \delta_i \) must satisfy the following constraints:

Figure 14

\[ \delta_i > \hat{p}_i, \quad i = 0, 1 \]  

(12)

Figure 15

\[ \psi_i > \hat{q}_i, \quad i = 0, 1 \]  

(13)

and

Figure 16

\[ \delta_i + \psi_i > 1, \quad i = 0, 1 \]  

(14)

It can be easily shown that \( p_i^* \), of equation 9 (or \( q_i^* \), of equation 10) is admissible if \( \delta_i \) and \( \delta_i \), \( i = 0, 1 \), are feasible. Note that equations 12-14 are merely one set of feasibility constraints for \( \delta_i \) and \( \delta_i \), \( i = 0, 1 \) so that equations 9-10 are admissible estimators for the unknown \( p_i \) and \( q_i \). By simply reversing the direction of inequalities in equations 12-14, we could obtain another set of feasibility constraints. Mathematically, these two sets of feasibility constraints are equivalent because they are just mirror images of one another with respect to the straight line of \( \delta_i = 1 \) in the two dimensional space of ordered pairs \((\delta_i, \delta_i)\) \((\delta_i)\). But, equation 14 is preferable because it has a practical implication, that is, a good classification procedure should perform better than random \((\delta_i)\). In addition, the variance of equation 9 is readily given by

Figure 17

\[ \text{Var}(p_i^*) = \text{Var}(\hat{p}_i)/\Delta_i^2, \quad i = 0, 1 \]  

(15)

where \( \text{Var}(p_i) \) is given by equation 7.

By replacing the true unknown parameters \( p_i \) and \( q_i \), \( i = 0, 1 \), in equation 1 with equations 9-10, the bias-adjusted [exposure] odds ratio (BAOR) \( R^* \) is defined by

Figure 18

\[
R^* = \frac{p_i^* q_i^*}{p_i q_i} = \frac{(\psi_i + \hat{p}_i - 1)(\delta_i - \hat{p}_i)}{(\psi_i + \hat{p}_i - 1)(\delta_i - \hat{p}_i)}
\]  

(16)

where \( \delta_i \) and \( \delta_i \), \( i = 0, 1 \), are feasible. The BAOR of equation 16 is said to be admissible if it is a positive real number. Note that because of the feasibility constraints of equations 12-14, equation 16 is always a positive real number once true sensitivity and specificity are given. Clearly, equation 16 is admissible if \( p_i^* \), of equation 9 (or \( q_i^* \), of equation 10) are admissible.

By using the delta method, the asymptotic variance of \( \ln(R^*) \) is given by \((\tau)\)

Figure 19

\[ \text{Var}(\ln(R^*)) = \sum_{i=0}^1 \frac{\text{Var}(\hat{p}_i)}{(\Delta_i^2 p_i^* q_i^*)}, \]  

(17)

where \( p_i \) and \( q_i \), \( i = 0, 1 \), are given, respectively, by Table 1, \( \delta_i \) and \( \delta_i \), \( i = 0, 1 \), are feasible, and \( \text{Var}(p_i) \) is given by equation 15. A detailed derivation of equation 17 is given in the appendix. The approximation of the right side of equation 17 is adequate so long as \( n \) and \( n_1 \), sample sizes for cases and controls, are sufficiently large. When equation 17 is used in practical applications, the unknown parameters of \( p_i \) and \( q_i \), \( i = 0, 1 \), are replaced, respectively, by \( p_i^* \), of equation 9 and \( q_i^* \), of equation 10, and the true classification rates can be calculated exactly from the observed data in the main study under an assumption that the correct classified table is known as shown later in the next section.

According to the asymptotic theory of large sample distribution \((\tau)\), the sampling distribution of the following test statistic

Figure 20

\[ Z^* = \ln(R^*)/\text{s.e.}(\ln(R^*)) \]  

(18)

can be shown to follow a standard normal distribution,
Bias-adjusted exposure odds ratio for misclassified data

where $s.e.(\ln(R^{*)})$ denotes the standard error of $\ln(R)$, that is, a square root of equation 17. To account for misclassification errors in identifying any risk factor, equation 18 will be used to test whether the adjusted odds ratio of equation 3.8 is significant or not. In addition, equation 18 can be used to find the $100\% \times (1-\iota)$ confidence interval ($0 < \iota < 1$) for $R$ is given by

$$100\% \times (1-\alpha) \text{ CI for } R = \exp(\ln(R^*) + z_{\alpha/2} \cdot s.e.(\ln(R^*)))$$

where $z_{\alpha/2}$ is the $100 \times (1-\iota/2)$ percentile of the unit normal distribution.

To use equations 16-19 we need to know the true sensitivity and specificity for cases and controls. I will show below by using two practical examples how to calculate the true sensitivity and specificity from the data of the main-study. Basically, we need to know what the truly classified table is. This information is contained in the observed data of the main study. Even though we do not know exactly what the truly classified table is, our reverse thinking hints us that the truly classified table must be one of the reclassified tables from the observed one in the main study. Hence, we can obtain the truly classified table by assuming hypothetically that it is simply a table which is (either under- or over-) misclassified from the observed one by 1 subject, or 2 subjects, or … in the exposed category. Once we obtain the [hypothetically] true table, we’re thus able to calculate the sensitivity (or specificity) from the observed and this true table according to the following formula, that is,

$$\text{Sensitivity (or Specificity)} = 1 - \text{error rate} = 1 - \frac{|\text{true - observed}|}{|\text{true + observed}|}$$

RESULTS

In our first example validation data is not available, while a validation sample is available in the second example.

Example 1 In a study of deaths caused by landslides that occurred in the State of Chuuk, Federated States of Micronesia, a case-control design was used for the study. The U.S. Geological Survey reported 265 landslides of various sizes and 12 of these landslides caused 43 deaths and injured more than one hundred people. A case is defined as a person who died as a result of the landslides. Proxies are identified by the surviving villagers to provide information for decedents or persons in the control group who are too young to answer. A control group of 52 survivors were interviewed regarding their experience during the landslides, while only 40 proxies were interviewed regarding the circumstances of the death of their relatives or neighbors, because the study team was unable to find proxies for three of the victims ($\omega$). For an illustrative purpose we only take one table from their study regarding whether a person see natural warning signs (Table 3a). Here the exposure variable E is defined by the event that a person did not see any natural warning signs. By inspection of Table 3a, 95% ($= 37/39$) of people for cases did not see any natural warning signs, but only 52% ($= 27/52$) for controls did not either. Indeed, the crude odds ratio (COR) is obtained as 17.1 with $p < 5.3 \times 10^{-6}$ by Fisher’s exact test ($\chi^2$). Hence, based on this COR value, a tentative inference is drawn that whether seeing natural warning sign is a significant risk factor for the death caused by landslides.

However, since the data were collected through an interview survey from proxies or survivors, there was a possibility that misclassification might have occurred. Suppose that Table 3a is misclassified. To account for the misclassification bias, we have to use equation 16 to calculate the bias-adjusted exposure odds ratio. In this study no validation data were collected at all. Nevertheless, I’ll show first how to calculate the true sensitivity and specificity for cases. Evidently, it depends on our knowledge about what a truly classified table is. Even though we do not know exactly what the truly classified table is, we’re confident that it must be one of those 37 tables by under-misclassifying 1 subject or over-misclassifying 1, or 2, …, or 36 subjects in the category of “No signs” as shown in the first two columns and continued in the fifth and sixth columns of Table 3bi. Since I do not know exactly which one of these 37 possible scenarios is a truly classified table, I simply assume that each one of them is a desired correctly classified table and then calculate one by one the values of sensitivity and specificity accordingly. By using equation 20, the sensitivity and specificity pairs for cases and controls are given, respectively, in Tables 3b(i-ii). By taking the first entry in column 3 of Table 3bi as an example, we obtain that $\iota_i = 1 - |38 - 37|/(38+37) = 1 - 0.0133 = 0.9867$. However, after checking out if feasibility constraints [or equations 12-14] were satisfied, only the first four pairs of sensitivity and specificity were found to be feasible (highlighted in Table 3bi). Similarly, although there were altogether 50 possible truly classified tables for controls, only 34 pairs of sensitivity and specificity were found to be feasible for controls (Table 3bii).

To see in what direction the misclassification might bias the
Bias-adjusted exposure odds ratio for misclassified data

BAOR from the null-hypothesis value, I used all four feasible pairs of sensitivity and specificity for cases and only ten pairs, under- or over-misclassified up to five subjects in the same category of “No signs”, for controls to compute 40 (= 4×10) BAORs. The results of these 40 BAORs with its 95% confidence intervals (CI) are given in Tables 3c(i-iv). By inspection of 95% CI in Tables 3c(i-ii), the BAORs were overall significantly biased further away from the null value (R = 1) than the COR if just one subject was over-misclassified in the category of “No signs” for cases, and up to five persons were either under- or over-misclassified for controls (Tables 3cii), while they were significant and biased yet a little toward the null value than the COR if just one subject was under-misclassified in the category of “No signs” for cases and up to five persons were either under- or over-misclassified for controls (Tables 3ci). However, Tables 3c(iii-iv) painted a totally different picture, namely, the BAOR was overall biased away, yet not significant, from the null value, provided that more than one person was under-misclassified in the category of “No signs” for cases and up to up to five persons were under- or over-misclassified in the same category for controls.

Table 3(a): The survey data whether a person saw natural warning signs for cases and controls.

<table>
<thead>
<tr>
<th>Natural warnings</th>
<th>Case</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>No signs</td>
<td>37</td>
<td>27</td>
</tr>
<tr>
<td>Signs</td>
<td>2</td>
<td>25</td>
</tr>
<tr>
<td>Marginal total</td>
<td>39</td>
<td>52</td>
</tr>
</tbody>
</table>

Figure 23

Table 3b(i): All possible pairs of sensitivity and specificity for cases.

<table>
<thead>
<tr>
<th>Cases correctly classified [ ( \rho = 0.95, \eta = 0.05 ) ]</th>
</tr>
</thead>
<tbody>
<tr>
<td>No signs</td>
</tr>
<tr>
<td>----------</td>
</tr>
<tr>
<td>38 (+1)</td>
</tr>
<tr>
<td>36 (+1)</td>
</tr>
<tr>
<td>35 (+2)</td>
</tr>
<tr>
<td>34 (+3)</td>
</tr>
<tr>
<td>33 (+4)</td>
</tr>
<tr>
<td>32 (+5)</td>
</tr>
<tr>
<td>31 (+6)</td>
</tr>
<tr>
<td>30 (+7)</td>
</tr>
<tr>
<td>29 (+8)</td>
</tr>
<tr>
<td>28 (+9)</td>
</tr>
<tr>
<td>27 (+10)</td>
</tr>
<tr>
<td>26 (+11)</td>
</tr>
<tr>
<td>25 (+12)</td>
</tr>
<tr>
<td>24 (+13)</td>
</tr>
<tr>
<td>23 (+14)</td>
</tr>
<tr>
<td>22 (+15)</td>
</tr>
<tr>
<td>21 (+16)</td>
</tr>
<tr>
<td>20 (+17)</td>
</tr>
<tr>
<td>19 (+18)</td>
</tr>
</tbody>
</table>

Figure 24

Table 3b(ii): All feasible pairs of sensitivity and specificity for controls.
Bias-adjusted exposure odds ratio for misclassified data

Figure 25

<table>
<thead>
<tr>
<th>Controls (correctly classified)</th>
<th>( \psi_0 = 0.52, \psi_1 = 0.48 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>\text{No sign} ( ^a )</td>
<td>\text{Sign} ( \psi_0 )</td>
</tr>
<tr>
<td>44 (17) ( ^* )</td>
<td>8</td>
</tr>
<tr>
<td>43 (16)</td>
<td>9</td>
</tr>
<tr>
<td>42 (15)</td>
<td>10</td>
</tr>
<tr>
<td>41 (14)</td>
<td>11</td>
</tr>
<tr>
<td>40 (13)</td>
<td>12</td>
</tr>
<tr>
<td>39 (12)</td>
<td>13</td>
</tr>
<tr>
<td>38 (11)</td>
<td>14</td>
</tr>
<tr>
<td>37 (10)</td>
<td>15</td>
</tr>
<tr>
<td>36 (9)</td>
<td>16</td>
</tr>
<tr>
<td>35 (8)</td>
<td>17</td>
</tr>
<tr>
<td>34 (7)</td>
<td>18</td>
</tr>
<tr>
<td>33 (6)</td>
<td>19</td>
</tr>
<tr>
<td>32 (5)</td>
<td>20</td>
</tr>
<tr>
<td>31 (4)</td>
<td>21</td>
</tr>
<tr>
<td>30 (3)</td>
<td>22</td>
</tr>
<tr>
<td>29 (2)</td>
<td>23</td>
</tr>
<tr>
<td>28 (1)</td>
<td>24</td>
</tr>
</tbody>
</table>

\( ^* \) The “+” (or “-”) sign inside the parenthesis denotes the number of persons to be over- (or under-) misclassified.

Table 3c: Bias-adjusted exposure odds ratios (95% CI) for all 40 feasible pairs of sensitivity and specificity for cases and controls when:

Example 2 The data used here is taken from a case-control study on sudden infant death syndrome (SIDS) \( (1) \). Among women for whom only interview data were examined, 122 out of 564 case mothers and 101 out of 580 control mothers reported antibiotic use during pregnancy. By using equation 1, the obtained crude odds ratio 1.31 with a 95% CI: 0.98–1.76 (p-value = 0.07) is not statistically significant. In this study a second external validation sample based on the medical record (a gold standard) was available. From the data of this validation sample, the estimated sensitivity and specificity for cases and controls were given by \( \psi_{01} = (0.6304, 0.8667) \) and \( \psi_{10} = (0.5676, 0.9333) \), respectively. By assuming these estimated values as if they were true sensitivity and specificity probability for the study population given by Table 3, we are able to calculate the corresponding correctly cell counts (rounded to the largest integer) for Table 3 which should be: \( n_{11}^* = 56, n_{10}^* = 40, n_{01}^* = 338, \) and \( n_{00}^* = 419. \) Now since \( n_{11}^* + n_{00}^* = 394 \) and \( n_{10}^* + n_{01}^* = 459 \) are not the same as the marginal total in the table of the main study, we therefore convert proportionally these new marginal totals 394 and 459 to match, respectively, overall to that of the main-study table (564 and 580) for the entire table. We thus obtained: \( n_{11}^{**} = 80, n_{10}^{**} = 51, n_{01}^{**} = 484, \) and \( n_{00}^{**} = 529. \) It seems
Bias-adjusted exposure odds ratio for misclassified data

rational to assume that this new table is the truly classified table for the main-study. Then, we are able to calculate the true sensitivity and specificity for this SIDS study which are given by \((\phi_1, \psi_1) = (0.7921, 0.9546)\) and \((\phi, \psi) = (0.6711, 0.9504)\). By substituting the above values into equation 16, the BAOR is given by 1.18 with a 95% CI: 0.90–1.56 (p-value = 0.23) which is not statistically significant. The BAOR-value 1.18 is biased toward the null value more than the COR (1.31). Hence, there seems no association between the use of antibiotic during pregnancy and the SIDS.

However, suppose that the validation sample is reliable. Then, according to the above calculation, 42 and 50 women are over-misclassified in the antibiotic use during pregnancy for cases and controls, respectively. These numbers of misclassification on the antibiotic use seem quite big. I therefore conducted a sensitivity analysis by assuming arbitrarily (under- or over-) misclassified numbers ranging from 4 to 40 subjects in the category of the antibiotic use. First, I calculated the sensitivity and specificity for cases and controls, respectively (Table 4b). Next, I calculated the BAORs by using the 14 pairs of sensitivity and specificity given in Table 4b. By browsing the result in Table 4c, almost all 14 BAORs were significant except the last two values which used the last two pairs of sensitivity and specificity of Table 4b. The last two pairs of sensitivity and specificity in Table 4b correspond to that 34 and 40 women are over-misclassified in the category of “use” for both cases and controls. As a result, this implies that the BAORs will become significant and biased away from the null value if less than 34 women are over-misclassified in the category of “use” for both cases and controls in the main study. If this error of misclassification sounds more reasonable, then the inference drawn from using the validation data is clearly misleading.

Table 4a: The data of SIDS study of the exposure variable of interview response between cases and controls.

**Figure 27**

<table>
<thead>
<tr>
<th>Interview response</th>
<th>Cases</th>
<th>Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Use</td>
<td>122</td>
<td>101</td>
</tr>
<tr>
<td>No use</td>
<td>442</td>
<td>479</td>
</tr>
</tbody>
</table>

| Marginal total     | 564   | 580      |

Table 4b: Pairs of the [true] sensitivity and specificity for cases and controls.

**Figure 28**

* The “+” (or “-”) sign inside the parenthesis denotes the number of persons to be over- (or under-) misclassified.

Table 4c: Bias-adjusted exposure odds ratios with its p-value and 95% confidence interval.

**DISCUSSION**

Some observations are worthwhile for a discussion below:

A unique strength of this paper is that a simple, yet powerful, method is presented to calculate the true sensitivity and specificity based on the data available in the main study only, regardless whether the validation data is available or not. Hence, this new method can free
Bias-adjusted exposure odds ratio for misclassified data

Researchers in epidemiology from the no-validation-data scare (Example 1). Of course, we do not know exactly in example 1 what the truly classified table is. Nevertheless, we can get a general feeling about the misclassification effect on the estimation of the true exposure odds ratio. When the validation data are available, our method can not only identify the truly classified table, but also assist to assess the reliability of the validation data as I did in Example 2.

For both cases and controls, the sensitivity and specificity have to be calculated in pairs. This is resulted from a fact that the marginal totals are required to be fixed in case-control studies. As a consequence, we can not arbitrarily assign values to sensitivity and specificity separately as was done in the probabilistic/traditional sensitivity analysis (10) or in the simulation study (5).

The significance of the BAOR depends not only on its sheer magnitude, but also on its standard error (Table 4c(iii-iv)). Note that equation 17 is a nonlinear function of the true sensitivity and specificity parameters; hence its value can become very large too.

Even just misclassifying one subject the direction of the misclassification effect for the under- and over-misclassified scenario can be very different (Example 1).

In both of the above examples I could conduct an exhaustive sensitivity analysis, but chose not do it, because my intention was just for the sake of illustration.

APPENDIX

Let $\parallel$ denote the estimation errors between $n_i \cdot p_i^\circ$ and $n_i \cdot E(p_i^\circ)$, namely,

$$\gamma_i = n_i \cdot (\hat{p}_i - E(\hat{p}_i)) = n_i \cdot (\hat{p}_i - \Delta_i \cdot p_i + \psi_i - 1), i = 0, 1$$  \hspace{1cm} (A.1)

where $p_i$'s are defined in Table 1, $p_i^\circ$, $\parallel$, and $\parallel$ are given, respectively, by equations 2-3, 5-6, and 11. By using equations 2-3, equation A.1 can be also expressed in terms of $q_i^\circ$ as follows:

$$\gamma_i = n_i \cdot (p_i - p_i^\circ - \Delta_i \cdot p_i + \psi_i - 1), i = 0, 1$$  \hspace{1cm} (A.2)

It is easily shown that the first two moments of $\parallel$ are given, respectively, by

$$\gamma_i = n_i \cdot (p_i - p_i^\circ - \Delta_i \cdot p_i + \psi_i - 1), i = 0, 1$$  \hspace{1cm} (A.3)

and

$$\gamma_i = n_i \cdot (p_i - p_i^\circ - \Delta_i \cdot p_i + \psi_i - 1), i = 0, 1$$  \hspace{1cm} (A.4)

Equation 17 follows immediately from taking the variance of equation A.7 with the use of equations A.3-A.4 and independence property of $\parallel$, i = 0, 1.

ACKNOWLEDGEMENTS

This research was motivated by the author’s participation in the project of the Chuuk Landslide Study. The author would like to thank Dr. J. Malilay for her invitation to take part in that study.

Part of the results in this paper was presented by the author in the 2004 ASA Joint Statistical Meetings held in Toronto, Canada and subsequently being published in the ASA Proceedings (28).

CORRESPONDENCE TO

Tze-San Lee, Centers for Disease Control and Prevention, Mail Stop F-58, 4771 Buford Highway, Chamblee, GA 30341-3717, USA Phone: +1-770-488-3729; Fax: +1-770-488-1540 E-mail: tjls@cdc.gov

References

41. Willet W. An overview of issues related to the correction of non-differential exposure measurement error in epidemiologic studies. Stat
Bias-adjusted exposure odds ratio for misclassified data

Author Information

Tze-San Lee
Centers for Disease Control and Prevention