Available online at www.isas.org.in/jisas

**Journal of the Indian Society of Agricultural Statistics 67(2) 2013 183-195**

# The Effect on the Interaction from the Joint Misclassification of Two Exposure Factors in Contingency Tables

**Tze-San Lee**

*Western Illinois University, Centers for Disease Control and Prevention, Mail Stop F–60, Chamblee*

Received 10 September 2012; Accepted 29 January 2013

**SUMMARY**

In this paper we address the effect on the interaction from the joint misclassification of two exposure variables in three-way contingency tables. Two types of interaction, additive and multiplicative, are used to measure the effect of misclassification. Bias-adjusted cell proportions that account for the misclassification bias are presented. The data set of the lung cancer deaths from the mesothelioma tumors is used as an example to illustrate the effect on workers who are jointly exposed to two types of asbestos fibers, amphibole and chrysotile. Because no validation data are available, the theory of counterfactual is used to construct potential true [counterfactual] tables from the misclassified observed [factual] table. Because there are various possible true [counterfactual] tables, a study on how sensitive the effect on the interaction exerted by two misclassified exposure factors is then conducted. The result of the sensitivity analysis shows that the effect on the interaction by the joint misclassification of two exposure factors shouldn't be ignored. In particular, the inference of no interaction could be dramatically changed under either the additive or multiplicative criterion if the data are misclassified.

*Keywords:* Additive/ multiplicative interaction, Asbestos, Counterfactuals, Misclassification, Mesothelioma cancers.

## 1. INTRODUCTION

In epidemiological research some earlier authors were merely interested in what the joint effect from two exposure factors on the disease outcome is, while some recent authors were more interested in determining if the interaction effect between the two exposure factors is synergistic. On the one hand, Cornfield (1962) investigated the joint dependence on coronary heart disease of serum cholesterol and systolic blood pressure; Vincent and Marchetta (1963) and Keller and Terris (1965) studied the joint effect of alcohol and tobacco on the cancer of mouth, pharynx or larynx; Saracci (1977) investigated the epidemiologic evidence on the asbestos-smoking interaction on the lung cancer. In addition, see works also done by Berry *et al.* (1972), Rothman and Keller (1972), Ottman (1996), Lee (2001), Okamoto and Horisawa (2007), and Boffetta *et al.* (2012). On the other hand, in 1974 Rothman introduced the concept of synergy and antagonism in studying the cause-effect relationship and then in 1976 proposed a statistical method to evaluate the synergy and antagonism in order to complete the description of cause-effect relationship; Kupper and Hogan (1978) provided a review on the notion of interaction for a quantification of the joint effect of two or more potential risk factors acting in combination. However, in observational studies, it is highly plausible that two exposure factors are simultaneously misclassified. To provide an illustration, we use a study on the combined effect of two types of asbestos on the mesothelioma lung cancer (Gardener and Munford 1980). The specimens of asbestos were classified according to the number of asbestos fibers seen per grid and the type of fiber in the lung for the mesothelioma patients to distinguish chrysotile and amphibole (Pooley 1976). Evidently, such a procedure of classification might involve a possible misclassification of these two types

*E-mail address* : tjl3@cdc.gov

of asbestos fibers. Indeed, a study of the possible effect of such a joint misclassification of asbestos types on the mesothelioma in a case-control study will be presented in this paper.

In fact, misclassification is a common problem in epidemiological studies and a considerable amount of works on the deleterious effect of misclassification on a single exposure variable in case-control studies has been studied by various authors (Cochran 1968, Copeland *et al*. 1977, Fleagal *et al*. 1986, Fleiss *et al*. 2003, Kuha *et al*. 2001, Morrissey and Spiegelman 1999, Rothman and Greenland 1998, Thomas *et al*. 1993, Walter and Irwing 1988). Usually, only the exposure variable is assumed to be misclassified, but the outcome variable is measured correctly (Lagakos 1988). Some works are also published on the joint misclassification of the exposure and the outcome variable. For instance, Keys and Kihlberg (1963) studied the case that both the disease variable and the exposure variable are simultaneously subject to misclassification, while Kristensen (1992) investigated the scenario in which the bias comes from that the joint misclassification between exposure and outcome is non-differential yet dependent. Brenner *et al*. (1993) investigated also the effect of joint misclassification of exposure and disease on cumulative incidence ratio in the context of cohort studies.

Kleinbaum *et al*. (1982) studied the effect of misclassification under the assumption that the relationship between the exposure and the outcome variables is independent. Although Barron (1977) investigated two jointly misclassified random variables, he made an assumption of independence between the two variables.

Apparently, Tzonou *et al*. (1986) studied the effect of joint misclassification of two dichotomous risk factors in case-control studies. But, what they actually studied was the misclassification of an exposure variable and a confounding variable. Chiacchierini and Arnold (1977) investigated a problem in which $2 \times 2$ contingency tables with both margins are subject to misclassification. But their study was limited to the case in which among two methods of classification, one method was error-free whereas the other was fallible. Fung and Howe (1984) investigated the effect of joint misclassification between a multiple-level risk factor and a confounding factor on the estimation of the relative risk and statistical power in case-control studies. More recently, Garcia-Closas *et al*. (1999) considered the misclassification effect of gene-environment interaction on the assessment of bias and sample size requirement. Tarafder *et al*. (2011) assessed the impact of misclassification error on the association of between the soil-transmitted helminth and the incidence of schistosoma japonium infection. But, no works has ever dealt with the misclassification of two exposure factors simultaneously.

In this paper the issue of joint misclassification from two binary exposure variables is addressed in the context of a case-control study. The joint misclassification error probabilities are first defined. Bias-adjusted estimates for cell proportions of four cell probabilities are then presented when two exposure factors are jointly misclassified. Two types of interaction criteria, additive and multiplicative, are used to measure the effect of misclassification. Asymptotic standard errors for the bias-adjusted estimates (or its logarithm) are derived. The data set for the lung cancer of mesotheliomas is used to illustrate on how to calculate the true misclassification probabilities by using the theory of counterfactuals if the validation data are not available. Because there are various possible true [counterfactual] tables, a sensitivity analysis is carried out to study the effect on the interaction from the joint misclassification of workers who are exposed to two different types of asbestos fibers,

## 2. BACKGROUND

Consider a case-control study with two exposure factors. Let $E$ and $F$ be two binary exposure variables (1, if yes; 0, otherwise), and $D$ denote the disease variable (1 if cases, 0 if controls). Suppose that under the multinomial sampling design a simple random sample of sizes $n_{[1]}$ and $n_{[0]}$ are selected with respect to both $E$ and $F$ from the case and control groups, respectively. Furthermore, assume that instead of being classified by $E$ and $F$ the collected data are classified jointly by $E^*$ and $F^*$ which are surrogate variables for $E$ and $F$, respectively. Let $n_{ijk}$ denote the collected data as shown in Table 1 with the index $i, j, k$ corresponding, respectively, to the variables $E$, $F$ and $D$. As a result, for fixed $k$ $\{n_{ijk}\}$ is assumed to follow a multinomial distribution with cell probabilities $\pi_{ijk}$ and the fixed sample size $n_{[k]}$.

**Table 1.** Observed cell frequencies in a 2 × 2 × 2 table which are classified by two surrogate exposure factors $E^*$ and $F^*$

| | $D$ = 1 (Cases) | | | $D$ = 0 (Controls) | |
|---|---|---|---|---|---|
| | $F^*$= 1 | $F^*$= 0 | | $F^*$= 1 | $F^*$= 0 |
| $E^*$ = 1 | $n_{111}$ | $n_{101}$ | $E^*$ = 1 | $n_{110}$ | $n_{100}$ |
| $E^*$ = 0 | $n_{011}$ | $n_{001}$ | $E^*$ = 0 | $n_{010}$ | $n_{000}$ |
| Sample size | $n_{[1]}$ | | Sample size | $n_{[0]}$ | |

**Note.** In Table 1, $n_{[k]} = \sum_{i,j=0}^{1} n_{ijk}$ for $k$ = 0, 1.

For all possible classifications of $E^*$and $F^*$ in the actual study population, we define all 16 possible jointly classified conditional probabilities of $E^*$ and $F^*$, conditioned on that the true $E$ and $F$ are known, as follows: for fixed $k$ (= 0 or 1) and $i'$, $j'$, $i, j$ = 0, 1

$$\delta^{[ij]}_{i'j'(k)} = \Pr(E^* = i,\ F^* = j \mid E = i',\ F = j';\ D = k) \quad (1)$$

where $\{\delta^{[ij]}_{i'j'(k)}\}$, for $i', j'$ = 0, 1, are required to satisfy the following identities:

$$\sum_{i,j=0}^{1} \delta^{[ij]}_{i'j'(k)} = 1,\ 0 \le \delta^{[ij]}_{i'j'(k)} \le 1. \quad (2)$$

Because of the constraints of equation (2), only 12 out of a total of 16 conditional probabilities of Equation (1) are regarded as free parameters for fixed $k$. Throughout this paper, we will use those 12 misclassification probabilities in equation 1 as free parameters for either cases or controls separately. The misclassification is said to be nondifferential if these 12 misclassification probabilities are equal one another for case and control groups; it is said to be differential otherwise. In this paper only the differential misclassification is considered. Furthermore, assume that there is no confounding or selection bias hidden in the collected data. If no misclassification is thought to be present in either $E^*$ or $F^*$, the observed sample proportions, for fixed $k$, given by

$$\hat{\pi}_{ijk} = n_{ijk} / n_{[k]},\ i, j = 0, 1, \quad (3)$$

are shown to be unbiased estimators for the true unknown cell probabilities $\pi_{ijk}$ in Table 1 and the variance-covariance matrix $\Sigma_{[k]}$ of $\hat{\pi}_{[k]}$ is given by (Agresti 2002).

$$\Sigma_{[k]} \equiv [\sigma_{ijk}]_{i,j=1}^{4}$$

$$= n_{[k]}^{-1}\begin{bmatrix} \pi_{11k}\bar{\pi}_{11k} & -\pi_{11k}\pi_{10k} & -\pi_{11k}\pi_{01k} & -\pi_{11k}\pi_{00k} \\ -\pi_{11k}\pi_{10k} & \pi_{10k}\bar{\pi}_{10k} & -\pi_{10k}\pi_{01k} & -\pi_{10k}\pi_{00k} \\ -\pi_{11k}\pi_{01k} & -\pi_{10k}\pi_{01k} & \pi_{01k}\bar{\pi}_{01k} & -\pi_{01k}\pi_{00k} \\ -\pi_{11k}\pi_{00k} & -\pi_{10k}\pi_{00k} & -\pi_{01k}\pi_{00k} & \pi_{00k}\bar{\pi}_{00k} \end{bmatrix} \quad (4)$$

where $\bar{\pi}_{ijk} = 1 - \pi_{ijk}$.

However, if $E^*$ and $F^*$ are misclassified, equation (3) is no longer unbiased and its expected values are corrected by for fixed $k$,

$$\mathrm{E}(\bar{\pi}_{[k]}) = W_{[k]}\pi_{[k]}, \quad (5)$$

where column vectors $\hat{\pi}_{[k]}$, $\pi_{[k]}$, and the matrix $W_{[k]}$ are given, respectively, by

$$\hat{\pi}_{[k]} = (\hat{\pi}_{11k}, \hat{\pi}_{10k}, \hat{\pi}_{01k}, \hat{\pi}_{00k})^T \quad (6)$$

$$\pi_{[k]} = (\pi_{11k}, \pi_{10k}, \pi_{01k}, \pi_{00k})^T \quad (7)$$

$$W_{[k]} = \begin{bmatrix} 1-\delta^{[10]}_{11(k)} & \delta^{[11]}_{10(k)} & \delta^{[11]}_{01(k)} & \delta^{[11]}_{00(k)} \\ -\delta^{[01]}_{11(k)} & -\delta^{[00]}_{11(k)} & & \\ \delta^{[10]}_{11(k)} & 1-\delta^{[11]}_{10(k)} & \delta^{[10]}_{01(k)} & \delta^{[10]}_{00(k)} \\ & -\delta^{[01]}_{10(k)} & -\delta^{[00]}_{10(k)} & \\ \delta^{[01]}_{11(k)} & \delta^{[01]}_{10(k)} & 1-\delta^{[11]}_{01(k)} & \delta^{[01]}_{00(k)} \\ & & -\delta^{[10]}_{01(k)} & -\delta^{[00]}_{01(k)} \\ \delta^{[00]}_{11(k)} & \delta^{[00]}_{10(k)} & \delta^{[00]}_{01(k)} & 1-\delta^{[11]}_{00(k)} \\ & & & -\delta^{[10]}_{00(k)} \quad -\delta^{[01]}_{00(k)} \end{bmatrix} \quad (8)$$

and the superscript "$T$" in equations (7) and (8) denotes the transpose of a vector/matrix. By replacing the left-hand side of equation (5) with equation (3) and solve for the unknown $\pi_{[k]}$, the bias-adjusted cell proportion (BACP) estimator $\breve{\pi}_{[k]}$, which accounts for the misclassification bias for the true unknown $\pi_{ijk}$, is defined by

$$\breve{\pi}_{[k]} = W_{[k]}^{-1}\hat{\pi}_{[k]}, \quad (9)$$

where $W_{[k]}^{-1}$ is the inverse of $W_{[k]}$, and the column vectors $\bar{\pi}_{[k]}$ is given by

$$\bar{\pi}_{[k]} = (\bar{\pi}_{11k}, \bar{\pi}_{10k}, \bar{\pi}_{01k}, \bar{\pi}_{00k})^T .$$

In addition, the bias-adjusted cell count estimator (BACC) is defined by $\bar{n}_{[k]} = n_{[k]}.\bar{\pi}_{[k]}$, where $n_{[k]}$ is the sample size defined in equation (3) and $\bar{\pi}_{[k]}$ is given by equation (9). Note that equation (9) depends on the misclassification probabilities $\{\delta_{i'j'(k)}^{[ij]}\}$, for $i', j' = 0, 1$ which can be calculated through applying the theory of counterfactuals to the observed (or misclassified) table available in the main study. The details of calculation will be illustrated in the section of "Example".

Although it is possible to find a closed form formula for $W_{[k]}^{-1}$ by using the MATHEMATICA (Lee 2007), it is less messy to solve instead equation (9) numerically as a system of linear equation for $\bar{\pi}_{[k]}$. Incidentally, a set of misclassification probabilities $\{\delta_{i'j'(k)}^{[ij]}\}$, for $i', j' = 0, 1$, is said to be "feasible" if all $\{\delta_{i'j'(k)}^{[ij]}\}$, for $i', j' = 0, 1$ are numbers between 0 and 1 satisfying equation (2) and $\det(W_{[k]})$, the determinant of $W_{[k]}$, is non-zero for the given set of $\delta_{i'j'(k)}^{[ij]}$. Furthermore a set of feasible $\{\delta_{i'j'(k)}^{[ij]}\}$ for $i', j' = 0, 1$, is said to be "admissible" if for a set of feasible $\{\delta_{i'j'(k)}^{[ij]}\}$, we have $0 < \bar{\pi}_{ijk} < 1$ and $\sum_{j=0}^{1}\sum_{j=0}^{1} \bar{\pi}_{ijk} = 1$ for fixed $k$ and all $i, i', j, j' = 0, 1$.

## 3. METHOD

In a case-control study having two exposure variables, the conventional notion of odds ratio is no longer applicable (Birch 1964). Instead, we need to focus on the issue: "Whether these two exposure factors confer their risks independently or in some interactive fashion." Three possible definitions for interaction, one additive and two multiplicative, were considered by Darroch and Borkent (1994). Here I only adopted two of them, one is additive and the other one is multiplicative. Two exposure variables $E$ and $F$ are said to have no 2-way additive (or multiplicative) interaction if for fixed $k = 0, 1$

$$\theta_{addt\_2way}^{[k]} = 0 \quad (\text{or } \theta_{mult\_2way}^{[k]} = 1)$$

where $\theta_{addt\_2way}^{[k]}$ and $\theta_{mult\_2way}^{[k]}$ are defined respectively by

$$\theta_{addt\_2way}^{[k]} \equiv \pi_{11k} - \pi_{10k} - \pi_{01k} + \pi_{00k}, \tag{10}$$

$$\theta_{mult\_2way}^{[k]} \equiv \pi_{11k}\pi_{00k}(\pi_{10k}\pi_{01k})^{-1}. \tag{11}$$

For three variables $D$, $E$ and $F$, they are said to have no 3-way additive (or multiplicative) interaction (Bartlett 1935; Bhapkar and Koch 1968) if $\theta_{addt\_3way} = 0$ (or $\theta_{mult\_3way} = 1$), where

$$\theta_{addt\_3way} \equiv \theta_{addt\_2way}^{[1]} - \theta_{addt\_2way}^{[0]}$$

$$= \pi_{111} + \pi_{100} + \pi_{010} + \pi_{001} - \pi_{110} - \pi_{000} - \pi_{101} - \pi_{011} \tag{12}$$

$$\theta_{mult\_3way} \equiv \theta_{mult\_2way}^{[1]} / \theta_{mult\_2way}^{[0]}$$

$$= \pi_{111}\pi_{001}\pi_{100}\pi_{010}(\pi_{101}\pi_{011}\pi_{110}\pi_{000})^{-1}. \tag{13}$$

Clearly, the naive point estimators for $\theta_{addt\_2way}^{[k]}$, $\theta_{mult\_2way}^{[k]}$, $\theta_{addt\_3way}$ and $\theta_{mult\_3way}$ are given respectively, by

$$\hat{\theta}_{addt\_2way}^{[k]} = \hat{\pi}_{11k} - \hat{\pi}_{10k} - \hat{\pi}_{01k} + \hat{\pi}_{00k}, \tag{14a}$$

$$\hat{\theta}_{mult\_2way}^{[k]} = \hat{\pi}_{11k}\hat{\pi}_{00k}(\hat{\pi}_{10k}\hat{\pi}_{01k}), \tag{14b}$$

$$\hat{\theta}_{addt\_3way} \equiv \hat{\pi}_{111} + \hat{\pi}_{100} + \hat{\pi}_{010} + \hat{\pi}_{001} - \hat{\pi}_{110} - \hat{\pi}_{000} - \hat{\pi}_{101} - \hat{\pi}_{011}, \tag{14c}$$

$$\hat{\theta}_{mult\_3way} = (\hat{\pi}_{111}\hat{\pi}_{001}\hat{\pi}_{100}\hat{\pi}_{010})(\hat{\pi}_{101}\hat{\pi}_{011}\hat{\pi}_{110}\hat{\pi}_{000})^{-1} \tag{14d}$$

where $\{\hat{\pi}_{ijk}\}$, $i, j = 0, 1$, are given respectively by equation (3). The asymptotic standard error of equations (14a-d) are given, respectively, by

$$Var(\hat{\theta}_{addt\_2way}^{[k]}) = n_{[k]}^{-1}\{\sum_{i,j=1}^{0} \pi_{ijk}\bar{\pi}_{ijk} + 2(\pi_{11k}\pi_{10k} + \pi_{11K}\pi_{01k} - \pi_{11k}\pi_{00k} - \pi_{10k}\pi_{01k}$$

$$+\pi_{10k}\pi_{00k}+\pi_{01k}\pi_{00k})\} \quad (15a)$$

$$s.e.(\ln(\hat{\theta}^{[k]}_{mult\_2way})) = \sqrt{n^{-1}_{[k]}(\pi^{-1}_{11k}+\pi^{-1}_{10k}+\pi^{-1}_{01k}+\pi^{-1}_{00k})}, \quad (15b)$$

$$s.e.(\hat{\theta}_{addt\_3way}) = \sqrt{\sum_{k=0}^{1} Var(\theta^{[k]}_{addt\_2way})}, \quad (15c)$$

$$s.e.(\hat{\theta}_{mult\_3way})) = \sqrt{\sum_{k=0}^{1} n^{-1}_{[k]}(\pi^{-1}_{11k}+\pi^{-1}_{10k}+\pi^{-1}_{01k}+\pi^{-1}_{00k})} \quad (15d)$$

The details of derivation of equations (15a-d) are given in the appendix.

If both $E^*$ and $F^*$ are jointly misclassified, the bias-adjusted estimators for $\theta^{[k]}_{addt\_2way}$, $\theta^{[k]}_{mult\_2way}$, $\theta_{addt\_3way}$ and $\theta_{mult\_3way}$ are given respectively by

$$\breve{\theta}^{[k]}_{addt\_2way} = \breve{\pi}_{11k}-\breve{\pi}_{10k}-\breve{\pi}_{01k}+\breve{\pi}_{00k}, \quad (16a)$$

$$\breve{\theta}^{[k]}_{mult\_2way} = \breve{\pi}_{11k}\breve{\pi}_{00k}(\breve{\pi}_{10k}\breve{\pi}_{01k})^{-1}, \quad (16b)$$

$$\breve{\theta}_{addt\_3way} \equiv \breve{\pi}_{111}+\breve{\pi}_{100}+\breve{\pi}_{010}+\breve{\pi}_{001}-\breve{\pi}_{110} \breve{\pi}_{000}-\breve{\pi}_{101}-\breve{\pi}_{011}, \quad (16c)$$

$$\breve{\theta}_{mult\_3way} = (\breve{\pi}_{111}\breve{\pi}_{001}\breve{\pi}_{100}\breve{\pi}_{010})(\breve{\pi}_{101}\breve{\pi}_{011}\breve{\pi}_{110}\breve{\pi}_{000})^{-1} \quad (16d)$$

where $\{\breve{\pi}_{ijk}\}$, $i, j, k = 0, 1$, are given by equation 8. The BACP $\{\breve{\pi}_{ijk}\}$ (or BACC $\breve{n}_{ijk}$), $i, j, k = 0, 1$, of equation (8) are said to be plausible if for all admissible $\{\delta^{[ij]}_{i'j'(k)}\}$ we have $0 < \breve{\pi}_{ijk} < 1$ (or $\breve{n}_{ijk} > 0$) in which $\sum_{i,j=0}^{1} \breve{\pi}_{ijk} = 1$ (or $\sum_{i,j=0}^{1} \breve{n}_{ijk} = n_{[0]} + n_{[1]}$) for fixed $k$.

By conditioning on that the values of $\delta^{[ij]}_{i'j'(k)}$ are known and using Lemma 2.3.1 in Anderson (2003) or the delta method described in Chapter 14 of Agresti (2002), the asymptotic standard error of equations (16a-d) are given respectively by

$s.e.(\breve{\theta}^{[k]}_{addt\_2way})$

$$= \sqrt{\sum_{i=1}^{4} u_{iik} - 2(u_{12k}+u_{13k}+u_{24k}+u_{34k}-u_{14k}-u_{23k})} \quad (17a)$$

$$s.e.(\ln(\breve{\theta}^{[k]}_{mult\_2way})) = \sqrt{Var(\ln(\breve{\theta}^{[k]}_{\text{mult\_2way}})} \quad (17b)$$

$$s.e.(\breve{\theta}_{addt\_3way}) = \sqrt{\sum_{k=0}^{1}[n^{-1}_{[k]}Var(\breve{\theta}^{[k]}_{addt\_2way})]} \quad (17c)$$

$$s.e.(\ln(\breve{\theta}_{mult\_3way})) = \sqrt{\sum_{k=0}^{1}[Var(\ln(\breve{\theta}^{[k]}_{mult\_2way}))]}, \quad (17d)$$

where $\{u_{ijk}\}$ in equation (17a) and are given by equation (A9) in the Appendix, and the details for the derivation of equations (17a-d) are also given in the appendix. By the way, the value under the square root bracket in equations (17a-d) is shown to be positive in the Appendix.

Based on the asymptotic large sample theory, the sampling distribution for the test statistic for all the above estimates (equations (14a-d) or (16a-d)) can be shown to follow a standard normal distribution. Hence, the 100% × (1 – $\alpha$) confidence interval (CI) ($0 < \alpha < 1$) for any θ is given, for no multiplicative interaction, by

100% × (1 – $a$) CI for $\theta$ :

$$= [\exp(\ln(\breve{\theta}) \pm z_{a/2} \times s.e.(\ln(\breve{\theta}))], \quad (18)$$

(or 100% × (1 – $a$) CI for $\theta$ :

$$= [(\breve{\theta}) \pm z_{a/2} \times s.e.(\breve{\theta}))]$$

for no additive interaction)

where $z_{\alpha/2}$ is the $(\alpha/2)^{th}$ upper-tail percentile of the standard normal distribution.

## 4. EXAMPLE

The data taken from Gardner and Munford (1980) is used to illustrate the bias-adjusted method in the previous section (Table 2). Lung tissues from 120 cases dying of mesothelioma cancer and from 135 controls dying from other causes were examined under electron microscope. The lung samples were examined for the presence of two asbestos fiber types, chrysotile and amphibole. Here the random variable $D$ represents whether a subject died of mesothelioma (= 1 if yes, = 0 if the subject was a control), while $E$ (= 1 if yes, = 0

**Table 2.** Observed cell counts of the lung cancer of mesothelioma for cases and controls were classified according to types of asbestos fibers found in the lung tissue

| | Cases | | Controls | |
|---|---|---|---|---|
| | Amphibole | | | |
| | 1 | 0 | 1 | 0 |
| Chrysotile 1 | 47 | 27 | 6 | 35 |
| Chrysotile 0 | 37 | 9 | 24 | 70 |
| Sample size | 120 | | 135 | |

if no) and $F$ (= 1 if yes, = 0 if no) denote respectively the asbestos fiber of chrysotile amphibole that were found in the lung.

Under the assumption that there is no joint misclassification between $E^*$ and $F^*$ the crude point estimators were estimated by using equations (3), (14a-d), and (15a-d) as $\hat{\theta}^{[1]}_{addt\_2way} = -0.067$ (p = 0.23), $\hat{\theta}^{[1]}_{mult\_2way} = 0.42$ (p = 0.026) for cases, while $\hat{\theta}^{[0]}_{addt\_2way} = 0.126$ (p = 0.07), $\hat{\theta}^{[0]}_{mult\_2way} = 0.5$ (p = 0.08) for controls. Furthermore, $\hat{\theta}_{addt\_3way} = -0.193$ (p = 0.06) and $\hat{\theta}_{mult\_3way} = 0.85$ (p = 0.40). For the 2-way interaction, an inference is tentatively drawn from their p-values: a hypothesis that the interaction is additive is rejected neither for cases (p = 0.23) and nor for controls (p = 0.07); however, a hypothesis that the interaction is multiplicative is rejected for cases (p = 0.026), but not for controls (p = 0.08). For the 3-way interaction, a hypothesis that the interaction is is rejected neither for being multiplicative (p = 0.40), and nor for being additive (p = 0.06).

Suppose that "chrysotile" and "amphibole" in Table 2 are jointly misclassified for both cases and controls as illustrated in the section of introduction. Now the problem we're facing is to calculate the value of misclassification probability $\{\delta^{[ij]}_{i'j'(k)}\}$, for $i', j' = 0, 1$, under the restraint that we do not have the validation sample. Here counterfactual thinking comes into play (Epstude and Roese 2008), that is, if only we know the true contingency table, we're then able to calculate the value of misclassification probability from the observed table which is the misclassified one and the true table which serves as our "gold standard". Evidently, the potential true table, even though unknown, can be figured out from the observed table as shown below. Since we do not know which potential outcome table is the genuine true table, we're required to consider all possible outcome tables figuring out from the observed table. This eventually leads to the sensitivity analysis for all possible true tables. Note that although the observed table was misclassified, it was the factual one. Thus, even though the true table is unknown, it was nothing but a counterfactual one to the observed table. Consequently, the potential true table could be constructed from the observed table. The only thing new is that there are many potential counterfactual tables and we have no way to be certain that which one of these possible counterfactual tables is the genuine true table. Because of this difficulty, we are forced to incorporate a sensitivity analysis on all possible potential true tables into our study later.

Because the column/row marginal totals are required to be fixed in case-control studies, only one out of four cell frequencies can be regarded as a free parameter in either cases or controls. To construct true [counterfactual] tables, the frequency in cell (0, 0) (or cell (1, 1)) was adopted as a free parameter for cases (or controls). In 15 true [counterfactual] tables for cases, the frequency in cell (0, 0) took values of 3, 4, …, 8, 10, 11, …, 18. Frequencies in cells (1, 0) and (0, 1) were obtained by subtracting the frequency of cell (0, 0) from the corresponding column/row marginal total. The frequency in cell (1, 1) was obtained by subtracting frequencies in cell (1, 0), (0, 1), and (0, 0) from the grand total in the observed table (Table 3a, column 2). Another 15 true [counterfactual] tables were similarly constructed for controls (Table 3b, column 2). A classification procedure is said to be under- (or over-) misclassified if the designated cell in the true [counterfactual] table has its frequency more (or less) than the observed [factual] table. For convenience, we put the least under- and/or over-misclassified true [counterfactual] tables in the middle of the first column. For cases models #6 and #7 were the least over- and under-misclassified, respectively, merely by one subject in cell (0, 0), so do models #5 and #6 in cell (1, 1) for controls. The amounts of misclassification were increased gradually in the remaining models.

Now, by conditioning on that the true table is known, we're able to calculate the corresponding

$\{\delta_{ij'(k)}^{[ij]}\}$, for $i', j' = 0, 1$. Before starting to compute the misclassification probabilities, note that for each cell frequency there are three possible ways to misclassify cell frequencies between the true [counterfactual] table and the misclassified observed [factual] table. For example, the frequency in cell (1, 1) of the true table might be misclassified into the other three cells of the observed table, (1, 0), (0, 1) and (0, 0). Because all three ways of misclassification are equally likely to occur, we assume $\delta_{11(k)}^{[10]} = \delta_{11(k)}^{[01]} = \delta_{11(k)}^{[00]} = \frac{1}{3}(1-\delta_{11k}^{[11]})$ for cell (1, 1). Similar equations were assumed to hold for other cells too.

By using the observed cell frequencies in Table 2, we first calculated the misclassification probabilities

**Table 3.** (a) Fifteen models (or true [counterfactual] tables) with its misclassification probabilities for cases.

| Model (#) | $(n_{111}, n_{101}, n_{011}, n_{001})$ | $(\delta_{11(1)}^{[11]}, \delta_{11(1)}^{[10]} = \delta_{11(1)}^{[01]} = \delta_{11(1)}^{[00]}, \delta_{10(1)}^{[10]}, \delta_{10(1)}^{[11]} = \delta_{10(1)}^{[01]} = \delta_{10(1)}^{[00]}, \delta_{01(1)}^{[01]}, \delta_{01(1)}^{[11]} = \delta_{01(1)}^{[10]} = \delta_{01(1)}^{[00]}, \delta_{00(1)}^{[00]}, \delta_{00(1)}^{[11]} = \delta_{00(1)}^{[10]} = \delta_{00(1)}^{[01]})$ | $\det(W_{[1]})$ |
|---|---|---|---|
| 1 | (41, 33, 43, 3) | (0.93, 0.02, 0.9, 0.03, 0.93, 0.03, 0.5, 0.17) | 0.38 |
| 2 | (42, 32, 42, 4) | (0.94, 0.02, 0.92, 0.03, 0.94, 0.02, 0.62, 0.13) | 0.49 |
| 3 | (43, 31, 41, 5) | (0.96, 0.01, 0.93, 0.02, 0.95, 0.02, 0.71, 0.1) | 0.60 |
| 4 | (44, 30, 40 6) | (0.97, 0.01, 0.95, 0.02, 0.93, 0.01, 0.8, 0.07) | 0.70 |
| 5 | (45, 29, 39, 7) | (0.98, 0.01, 0.96, 0.01, 0.97, 0.01, 0.87, 0.04) | 0.80 |
| 6 | (46, 28, 38, 8) | (0.99, 0.004, 0.98, 0.01, 0.99, 0.004, 0.94, 0.02) | 0.90 |
| 7 | (48, 26, 36, 10) | (0.99, 0.004, 0.98, 0.01, 0.99, 0.005, 0.95, 0.02) | 0.91 |
| 8 | (49, 25, 35, 11) | (0.98, 0.01, 0.96, 0.01, 0.97, 0.01, 0.9, 0.03) | 0.82 |
| 9 | (50, 24, 34, 12) | (0.97, 0.01, 0.94, 0.02, 0.96, 0.01, 0.86, 0.05) | 0.75 |
| 10 | (51, 23, 33, 13) | (0.96, 0.01, 0.92, 0.03, 0.94, 0.02, 0.82, 0.06) | 0.68 |
| 11 | (52, 22, 32, 14) | (0.95, 0.02, 0.90, 0.03, 0.93, 0.02, 0.78, 0.07) | 0.61 |
| 12 | (53, 21, 31, 15) | (0.94, 0.02, 0.87, 0.04, 0.91, 0.03, 0.75, 0.08) | 0.55 |
| 13 | (54, 20, 30, 16) | (0.93, 0.02, 0.85, 0.05, 0.9, 0.03, 0.72, 0.09) | 0.50 |
| 14 | (55, 19, 29, 17) | (0.92, 0.03, 0.83, 0.06, 0.88, 0.04, 0.69, 0.1) | 0.45 |
| 15 | (56, 18, 28, 18) | (0.91, 0.03, 0.8, 0.07, 0.86, 0.05, 0.67, 0.11) | 0.41 |

(b) Fifteen models (or true [counterfactual] tables) with its misclassification probabilities for controls

| # | $(n_{110}, n_{100}, n_{010}, n_{000})$ | $(\delta_{11(0)}^{[11]}, \delta_{11(0)}^{[10]} = \delta_{11(0)}^{[01]} = \delta_{11(0)}^{[00]}, \delta_{10(0)}^{[10]}, \delta_{10(0)}^{[11]} = \delta_{10(0)}^{[01]} = \delta_{10(0)}^{[00]}, \delta_{01(0)}^{[01]}, \delta_{01(0)}^{[11]} = \delta_{01(0)}^{[10]} = \delta_{01(0)}^{[00]}, \delta_{00(0)}^{[00]}, \delta_{00(0)}^{[11]} = \delta_{00(0)}^{[10]} = \delta_{00(0)}^{[01]})$ | $\det(W_{[1]})$ |
|---|---|---|---|
| 1 | (1, 40, 29, 65) | (0.29, 0.24, 0.93, 0.02, 0.91, 0.03, 0.96, 0.01) | 0.22 |
| 2 | (2, 39, 28, 66) | (0.50, 0.17, 0.95, 0.02, 0.92, 0.03, 0.97, 0.01) | 0.42 |
| 3 | (3, 38, 27, 67) | (0.67, 0.11, 0.96, 0.01, 0.94, 0.02, 0.98, 0.01) | 0.58 |
| 4 | (4, 37, 26, 68) | (0.80, 0.07, 0.97, 0.01, 0.96, 0.01, 0.986, 0.005) | 0.73 |
| 5 | (5, 36, 25, 69) | (0.91, 0.03, 0.986, 0.005, 0.98, 0.007, 0.993, 0.002) | 0.87 |
| 6 | (7, 34, 23, 71) | (0.92, 0.03, 0.986, 0.005, 0.98, 0.007, 0.993, 0.002) | 0.88 |
| 7 | (8, 33, 22, 72) | (0.86, 0.05, 0.97, 0.01, 0.96, 0.01, 0.986, 0.005) | 0.78 |
| 8 | (9, 32, 21, 73) | (0.80, 0.07, 0.96, 0.01, 0.93, 0.02, 0.979, 0.007) | 0.70 |
| 9 | (10, 31, 20, 74) | (0.75, 0.08, 0.94, 0.02, 0.91, 0.03, 0.97, 0.01) | 0.62 |
| 10 | (11, 30, 19, 75) | (0.71, 0.10, 0.92, 0.03, 0.88, 0.04, 0.97, 0.01) | 0.55 |
| 11 | (12, 29, 18, 76) | (0.67, 0.11, 0.91, 0.03, 0.86, 0.05, 0.96, 0.01) | 0.49 |
| 12 | (13, 28, 17, 77) | (0.63, 0.12, 0.89, 0.04, 0.83, 0.06, 0.95, 0.02) | 0.43 |
| 13 | (14, 27, 16, 78) | (0.60, 0.13, 0.87, 0.04, 0.80, 0.07, 0.95, 0.02) | 0.38 |
| 14 | (15, 26, 15, 79) | (0.57, 0.14, 0.85, 0.05, 0.77, 0.08, 0.94, 0.02) | 0.33 |
| 15 | (16, 25, 14, 80) | (0.55, 0.15, 0.83, 0.06, 0.74, 0.09, 0.93, 0.02) | 0.29 |

accordingly for each true [counterfactual] table and then used equation (2) to get the correct classification probabilities. The way to compute the misclassification probability is exactly the same as described in Lee (2009). Let's take $\delta_{10(1)}^{[11]}$ for Model 1 in Table 3a as an example. This meant that the frequency in cell (1, 0) of Model #1 was supposed to be 33, but because of under-misclassification the frequency in cell (1, 0) of the observed table was 27. As a result, six subjects in cell (1, 0) were under-misclassified. Out of the total 60 (= 33 + 27) subjects, the misclassification probability for cell (1, 0) was calculated as 0.1 (= 6/60). Since all three ways of misclassification were equally possible to occur, we thus obtained . $\delta_{10(1)}^{[11]} = \delta_{10(1)}^{[01]} = \delta_{10(1)}^{[00]} = \frac{0.1}{3}$ = 0.0333 ≈ 0.03. Note that all 15 true [counterfactual] tables (Table 3a, column 2) were feasible, because all misclassification probabilities plus their corresponding determinants were positive (Table 3a, columns 3-4). Similarly, all 15 true [counterfactual] tables for controls were found to be feasible too (Table 3b, columns 2).

For k = 0 and 1, we next proceeded to calculate the BACC $\breve{n}_{ij(k)}$. Thus, for all 15 feasible $\{\delta_{i'j'(k)}^{[ij]}\}$, for $i'$, $j'$ = 0, 1 in Tables 3a-b, we computed numerically the values of $\{\breve{n}_{ij(k)}\}$ from equation (9) by using the

**Table 4.** A check on the admissibility of MPs for 15 true [counterfactual] tables for cases and controls

| # | Controls: ($\breve{n}_{111}$, $\breve{n}_{101}$, $\breve{n}_{011}$, $\breve{n}_{001}$) | Controls: ($\breve{n}_{110}$, $\breve{n}_{100}$, $\breve{n}_{010}$, $\breve{n}_{000}$) |
|---|---|---|
| 1 | (46.3, 25.5, 35.7, 12.4) | (13.2, 32.5, 21.3, 89.5) |
| 2 | (46.7, 26.2, 36.3, 10.8) | (8.2, 34.2, 23.1, 69.5) |
| 3 | (46.9, 26.6, 36.6, 9.9) | (6.8, 34.7, 23.6, 69.8) |
| 4 | (47.0, 26.8, 36.8, 9.4) | (6.3, 34.9, 23.9, 69.9) |
| 5 | (47.0, 26.9, 36.9, 9.2) | (6.1, 35.0, 24.0, 70.0) |
| 6 | (47.0, 27.0, 37.0, 9.0) | (6.0, 35.0, 24.0, 70.0) |
| 7 | (47.0, 27.0, 37.0, 9.0) | (5.8, 35.1, 24.1, 70.0) |
| 8 | (47.0, 27.1, 37.1, 8.9) | (5.6, 35.2, 24.2, 70.0) |
| 9 | (47.0, 27.2, 37.1, 8.7) | (5.2, 35.3, 24.4, 70.1) |
| 10 | (47.0, 27.3, 37.2, 8.5) | (4.7, 35.5, 24.7, 70.1) |
| 11 | (46.9, 27.5, 37.4, 8.1) | (4.1, 35.7, 25.0, 70.1) |
| 12 | (46.9, 27.8, 37.6, 7.7) | (3.3, 36.0, 25.5, 70.2) |
| 13 | (46.9, 28.1, 37.8, 7.2) | (2.4, 36.4, 26.1, 70.2) |
| 14 | (46.8, 28.5, 38.1, 6.6) | (1.3, 36.8, 26.8, 70.2) |
| 15 | (46.7, 29.0, 38.4, 5.9) | (–0.1, 37.2, 27.7, 70.2) |

**Table 5.** Admissible bias-adjusted estimates (or equation (16)) for 2- or 3-way additive/multiplicative no-interaction with its p-value

(a) Bias-adjusted estimates for 2-way additive/ multiplicative interaction

| Model | $\breve{\theta}_{addt\_2way}^{[1]}$ (p-value) | $\breve{\theta}_{mult\_2way}^{[1]}$ (p-value) | $\breve{\theta}_{addt\_2way}^{[0]}$ (p-value) | $\breve{\theta}_{mult\_2way}^{[0]}$ (p-value) |
|---|---|---|---|---|
| 1 | –0.02 (0.44) | 0.63 (0.21) | 0.20 (0.09) | 1.30 (0.38) |
| 2 | –0.04 (0.36) | 0.53 (0.12) | 0.15 (0.07) | 0.72 (0.34) |
| 3 | –0.05 (0.31) | 0.48 (0.11) | 0.14 (0.07) | 0.58 (0.28) |
| 4 | –0.06 (0.27) | 0.45 (0.06) | 0.13 (0.07) | 0.53 (0.27) |
| 5 | –0.06 (0.27) | 0.44 (0.05) | 0.13 (0.068) | 0.51 (0.25) |
| 6 | –0.07 (0.23) | 0.42 (0.03) | 0.13 (0.068) | 0.50 (0.25) |
| 7 | –0.07 (0.23) | 0.42 (0.03) | 0.12 (0.089) | 0.48 (0.26) |
| 8 | –0.07 (0.23) | 0.42 (0.04) | 0.12 (0.094) | 0.46 (0.28) |
| 9 | –0.07 (0.24) | 0.41 (0.05) | 0.12 (0.10) | 0.42 (0.30) |
| 10 | –0.08 (0.21) | 0.39 (0.05) | 0.11 (0.13) | 0.38 (0.35) |
| 11 | –0.08 (0.22) | 0.37 (0.06) | 0.10 (0.16) | 0.32 (0.43) |
| 12 | –0.09 (0.19) | 0.35 (0.08) | 0.09 (0.19) | 0.25 (0.21) |
| 13 | –0.10 (0.17) | 0.32 (0.11) | 0.07 (0.25) | 0.18 (0.16) |
| 14 | –0.11 (0.16) | 0.28 (0.19) | 0.06 (0.29) | 0.09 (0.01) |
| 15 | –0.12 (0.14) | 0.25 (0.39) | * | * |

*not calculated because true [counterfactual] table #15 for controls is not admissible.

(b) Bias-adjusted estimates for 3-way additive/ multiplicative interaction

| | Model (Case) | | | |
|---|---|---|---|---|
| | $\breve{\theta}_{addt\_3way}$ (p-value) | | $\breve{\theta}_{mult\_3way}$ (p-value) | |
| Model (Control) | #1 | #15 | #1 | #15 |
| 1 | –0.22 (0.13) | –0.32 (0.04) | 0.48 (0.24) | 3.87 (0.07) |
| 2 | –0.17 (0.14) | –0.27 (0.04) | 0.88 (0.45) | 6.99 (0.01) |
| 3 | –0.16 (0.15) | –0.26 (0.04) | 1.09 (0.53) | 8.68 (0.01) |
| 4 | –0.15 (0.16) | –0.25 (0.04) | 1.19 (0.56) | 9.50 (0.02) |
| 5 | –0.15 (0.16) | –0.25 (0.04) | 1.24 (0.57) | 9.87 (0.01) |
| 6 | –0.15 (0.16) | –0.25 (0.04) | 1.26 (0.58) | 10.1 (0.02) |
| 7 | –0.14 (0.18) | –0.24 (0.05) | 1.31 (0.58) | 10.5 (0.02) |
| 8 | –0.14 (0.18) | –0.24 (0.05) | 1.37 (0.59) | 10.9 (0.04) |
| 9 | –0.14 (0.18) | –0.24 (0.05) | 1.50 (0.59) | 12.0 (0.07) |
| 10 | –0.13 (0.20) | –0.23 (0.06) | 1.66 (0.42) | 13.3 (0.15) |
| 11 | –0.12 (0.22) | –0.22 (0.07) | 1.97 (0.46) | 15.7 (0.34) |
| 12 | –0.11 (0.24) | –0.21 (0.08) | 2.52 (0.31) | 20.1 (0.04) |
| 13 | –0.09 (0.29) | –0.19 (0.10) | 3.50 (0.25) | 28.0 (0.03) |
| 14 | –0.08 (0.31) | –0.18 (0.12) | 7.00 (0.04) | 55.9 ($3.3\times10^{-5}$) |

software of MATLAB (2010) to find the determinant first of the misclassification matrix $W_{[k]}$ next, then the matrix inverse $W_{[k]}^{-1}$ and finally multiplied $W_{[k]}^{-1}$ by $\hat{n}_{[k]}$ to get the values of $\{\breve{n}_{ij(k)}\}$. According to the definition of admissibility of MPs, all models were admissible except model #15 for controls which was found to be inadmissible, because $\breve{n}_{110} = -0.1$ (Table 4, column 3). Since no validation sample data were available to determine which one out of 15 models was actually the true table, we therefore carried out the sensitivity analysis for all admissible models to examine the effect of joint misclassification from two exposure factors on the additive/multiplicative interaction parameters.

By using equations (16a-d)-(17a-d) all bias-adjusted estimates for 2- or 3-way additive/multiplicative interaction parameters were computed (Tables 5(a-b)). For both cases and controls all estimates for the 2-way additive interaction parameter are not significantly different from zero because their p-values are greater than 0.05 (Table 5(a), columns 2 & 4). Yet, there is a trend in the p-values. They are increasing from that of Models #5 and #6 in either direction of under-/over-misclassification for controls, but the pattern for cases is a little different. The p-values are increasing along the direction of the over-misclassification (from models #6 to #1), whereas decreasing along the direction of under-misclassification (from models #7 to #15). Yet, the pattern for estimates of the 2-way multiplicative interaction parameter are quite different from that for the additive scenarios. It depends on whether the group is cases or controls. For cases, the p-values are increasing along either direction of misclassification from 0.03 (model #6) to 0.21 (model #1) and again from 0.03 (model #7) to 0.39 (model #15). For controls, the pattern is somewhat different from that of cases. The p-values are increasing along the direction of over-misclassification from 0.25 (model #5) to 0.38 (model #1), but along the direction of under-misclassification it is increasing from 0.25 (model #6) to 0.43 (model #11) and then decreasing from 0.21 (model #12) to 0.01 (model #14).

For 3-way additive/multiplicative interaction parameter there are 210 possible models of combination from 15 models for cases and 14 models for controls. To calculate their estimates, we used all 14 models for controls, but only used two models (models #1 and #15) from cases. The p-values for all 3-way additive interaction estimates exhibit a pattern of increasing from 0.13 (model #1) to 0.31 (model #14) when case model = #1 and 0.04 to 0.12 when case model = #15 (Table 5(b), columns 2 and 3). For the 3-way multiplicative interaction estimates the p-values is decreasing both from 0.57 (model #5) to 0.24 (model #1) and from 0.58 (model #6) to 0.04 (model #14) when case model = #15 (Table 5(b), columns 4 and 5).

From the above limited sensitivity analysis it is easily seen that bias-adjusted estimates for 2- or 3-way multiplicative interaction parameter are more sensitive to whether the observed data are misclassified or not than for the additive interaction.

## 5. DISCUSSION

Some comments seem worthy to be given as follows:

1. By taking a quick glimpse at the data in Table 1, the inherent feature of the data for cases is intrinsically different from that for controls. The proportion of subjects in cases exposed to both types of asbestos is not low (0.39 =47/120), whereas that exposed to none of the two types of asbestos is low (0.08 = 9/120). In contrast, the proportion of subjects in controls exposed to both types of asbestos is low (0.04 = 6/135), whereas that exposed to none of the two types of asbestos is not low (0.52 = 70/135). For additional discussions on this data set, see Acheson and Gardner (1979).

2. Due to a small frequency ($n_{110} = 6$) in the (1, 1) cell for controls, it has two implications: (1) a small amount of misclassification results in large misclassification probabilities (Models #1-#3 in Table 3(b)); and (2) it has a much limited number of true [counterfactual] tables (it can have no more than six [over-misclassified] models as shown in Tables 3(b). It is likely attributed to this matter of large misclassification probability, model #15 is inadmissible, though feasible (Table 4).

3. Our 3-way additive interaction parameter (equation (12)) is different from that of the Gardner and Munford's (GM) (1980). This is because GM derived their 3-way additive

interaction parameter from using the notion of relative risk and the odds ratio was used by them to substituted for the relative risk under the rare disease assumption. Expressing in terms of our notations, it is given by

$$\theta_{addt\_GM} = \frac{\pi_{111}}{\pi_{110}} - \frac{\pi_{101}}{\pi_{100}} - \frac{\pi_{011}}{\pi_{010}} + \frac{\pi_{001}}{\pi_{000}}. \quad (19)$$

Because equation (19) is of the form of ratio of proportions, it is intuitively clear that the GM's 3-way additive interaction parameter must be very sensitive to the misclassification in Table 2. Nevertheless, the GM's 3-way multiplicative interaction parameter is exactly the same as equation (13). Under the no-misclassified-data assumption our result described in the second paragraph of Section 4 yields the same inference as that of the GM's, that is, the interaction for this data set is of being multiplicative rather than additive. Incidentally, we notice that the 2-way additive/multiplicative interaction parameters are unable to be defined by the GM's approach. Furthermore, our definitions do not need the rare disease assumption; hence the definition of interaction under additive or multiplicative is more universally applicable than the GM's.

4. Although definitions for additive/multiplicative interaction follow from the statistical stand-point rather than from the epidemiological perspective (Walter and Holford 1978), equations (10) and (11) do have an epidemiological interpretation. It defines an exact balance between the amounts of synergistic action and parallel action in the population who are exposed to both exposure factors, while equation (11) is a necessary, but not sufficient condition for the immunity model (Darroch and Borkent 1994).

5. As shown in Section 4, the inference for the 2-way multiplicative interaction for both cases and controls could drastically be changed depending how bad the data are misclassified. For example, the inference for cases can change from not multiplicative (models #6-#8) to multiplicative (models #1-#5 and #9-#14) and for controls the inference could change from multiplicative (models #1-#13) to not multiplicative (model #14) (Table 5(a), columns 3 and 5). Similarly, the inference for both the 3-way additive and multiplicative interaction estimate could be changed drastically depending how bad the data are misclassified (Table 5(b), columns 3 and 5).

6. Because the only unknown parameters are misclassification probabilities in equations 17(a-d) and we employed the theory of counterfactuals to calculate the misclassification probability exactly rather than estimating them, we have, therefore, no need to provide the estimated standard error. As a consequence, these formulas are not applicable to the case when the misclassification probabilities are estimated from the validation data once the validation sample dataset is available.

By the way, all numerical calculations done for Tables 3-5 were facilitated by using the Microsoft EXCEL spreadsheet and/or the software of MATLAB.

## 6. CONCLUSION

This paper presents a study on the effect of joint misclassification of two exposure factors on the interaction under two types of interaction, additive and multiplicative. Bias-adjusted cell proportions are presented to account for misclassification bias. The data taken from persons dying of the mesothelioma tumors were used as an example to illustrate the effect of being jointly exposed to asbestos fibers of amphibole and chrysotile. Since no validation data were available, the theory of counterfactual was employed to construct the true [counterfactual] table from the misclassified observed [factual] table. A sensitivity analysis was then conducted to see how sensitive the effect would be for various counterfactual true tables. From the result of the sensitivity analysis, it shows that the inference could be drastically changed depending how bad the data are misclassified.

Much research remains to be done in this area of the joint misclassification of two exposure factors. Just name a few: what happens if the two misclassified factors are polytomous variables? How to handle the issue of joint misclassification if the validation sample data are available? How does the misclassification affect the estimation of the attributable risk if two exposure factors are jointly misclassified?

## APPENDIX

Let the column vector c be defined as

$$c = [1, -1, -1, 1]^{T}. \quad \text{(A1)}$$

Thus, equation (14a) can be written in the form of

$$\hat{\theta}^{[k]}_{addt\_2way} = c^{T}\hat{\pi}_{[k]}, \quad \text{(A2)}$$

By using Lemma 2.3.1 in Anderson (2003), we then have

$$Var(\hat{\theta}_{addt\_2way}) = c^{T}\Sigma_{[k]}c$$

$$= \sum_{i=1}^{4}\sigma_{iik} - 2(\sigma_{12k} + \sigma_{13k} - \sigma_{14k} - \sigma_{23k} + \sigma_{24k} + \sigma_{34k}). \quad \text{(A3)}$$

Equation (15a) follows immediately from equations (A3) and (4). Because the variance-covariance matrix $\Sigma_{[k]^c}$ given by equation (4) is positive definite, the value of equation (A3) is clearly positive.

The derivation of equation 15b is exactly the same as equation (3.1) in Agresti (2002), p.71. By applying the independence between cases and controls to equation (15a), we have

$$Var(\hat{\theta}_{addt\_3way}) = \sum_{k=0}^{1} Var(\theta^{[k]}_{addt\_2way}). \quad \text{(A4)}$$

Equation (15c) follows directly from equations (A4) and (15a). Similarly, equation (15d) follows directly from applying the independence between cases and controls to equation (15b).

Let $v^{T}_{[i]k}$ be the $i^{\text{th}}$ row of the matrix $W^{-1}_{[k]}$. Thus, we have

$$\bar{\pi}_{[k]} = W^{-1}_{[k]}\hat{\pi}_{[k]}$$

$$= [v^{T}_{[1]k}\hat{\pi}_{[k]}, v^{T}_{[2]k}\hat{\pi}_{[k]}, v^{T}_{[3]k}\hat{\pi}_{[k]}, v^{T}_{[4]k}\hat{\pi}_{[k]}]^{T} \quad \text{(A5)}$$

Let the matrix $U$ be defined by

$$U = [u_{ijk}]^{4}_{i,j=1} \equiv W^{-1}_{[k]}\Sigma_{[k]}(W^{-1}_{[k]})^{T}. \quad \text{(A6)}$$

Again, by using Lemma 2.3.1 in Andersen (2003) on equation (A5), we have

$$Var(\bar{\theta}^{[k]}_{addt\_2way}) = c^{T}W^{-1}_{[k]}\Sigma_{[k]}(W^{-1}_{[k]})^{T}\,c$$

$$= c^{T}Uc \quad \text{(A7)}$$

Equation (17a) follows directly from equations (A1), (A6), and (A7). Since $\bar{\pi}_{[k]}$ is required to be admissible, this implies that the matrix $W^{-1}_{[k]}$ has to be positive. Consequently, the value of equation (A7) must be positive.

Let $g(\pi_{[k]}) = [\ln(v^{T}_{[1]k}\pi_{[k]}), \ln(v^{T}_{[2]k}\pi_{[k]}), \ln(v^{T}_{[3]k}\pi_{[k]}), \ln(v^{T}_{[4]k}\pi_{[k]})]^{T}$, where $\pi_{[k]} = [\pi_{11k}, \pi_{10k}, \pi_{01k}, \pi_{00k}]^{T}$. Thus,

$$\frac{\partial g}{\partial \pi_{[k]}} = (W^{-1}_{[k]})^{T}\,[diag(\pi^{*}_{[k]})]^{-1}, \quad \text{(A8)}$$

where $\pi^{*}_{[k]} = W^{-1}_{[k]}\pi_{[k]}$, and $diag(\pi^{*}_{[k]})$ is a diagonal matrix with the entries of $\pi^{*}_{[k]}$ as the diagonal entries.

Then, we have by using the delta method (or equation (14.8) in Agresti (2002))

$$Var(\ln(\bar{\theta}^{[k]}_{milt\_2way}) = n^{-1}_{[k]}\{c^{T}(W^{-1}_{[k]})^{T}\Sigma^{*}_{[k]}W^{-1}_{[k]}c\}, \quad \text{(A9)}$$

where $\Sigma^{*}_{[k]}$ is given by

$$\Sigma^{*}_{[k]} = [\sigma^{*}_{ij}]^{4}_{i,j=1}$$

$$= \begin{bmatrix} \frac{\pi_{11k}\bar{\pi}_{11k}}{\pi^{*2}_{11k}} & -\frac{\pi_{11k}\pi_{10k}}{\pi^{*}_{11k}\pi^{*}_{10k}} & -\frac{\pi_{11k}\pi_{01k}}{\pi^{*}_{11k}\pi^{*}_{01k}} & -\frac{\pi_{11k}\pi_{00k}}{\pi^{*}_{11k}\pi^{*}_{00k}} \\ -\frac{\pi_{11k}\pi_{10k}}{\pi^{*}_{11k}\pi^{*}_{10k}} & \frac{\pi_{10k}\bar{\pi}_{10k}}{\pi^{*2}_{10k}} & -\frac{\pi_{10k}\pi_{01k}}{\pi^{*}_{10k}\pi^{*}_{01k}} & -\frac{\pi_{10k}\pi_{00k}}{\pi^{*}_{10k}\pi^{*}_{00k}} \\ -\frac{\pi_{11k}\pi_{01k}}{\pi^{*}_{11k}\pi^{*}_{01k}} & -\frac{\pi_{10k}\pi_{01k}}{\pi^{*}_{10k}\pi^{*}_{01k}} & -\frac{\pi_{01k}\bar{\pi}_{01k}}{\pi^{*}_{01k}} & -\frac{\pi_{01k}\pi_{00k}}{\pi^{*}_{01k}\pi^{*}_{00k}} \\ -\frac{\pi_{11k}\pi_{00k}}{\pi^{*}_{11k}\pi^{*}_{00k}} & -\frac{\pi_{10k}\pi_{00k}}{\pi^{*}_{10k}\pi^{*}_{00k}} & -\frac{\pi_{01k}\pi_{00k}}{\pi^{*}_{01k}\pi^{*}_{00k}} & -\frac{\pi_{00k}\bar{\pi}_{00k}}{\pi^{*2}_{00k}} \end{bmatrix}$$

After a simplification of equation (A9), we have

$$Var(\ln(\bar{\theta}^{[k]}_{mult\_2way})$$

$$= n^{-1}_{[k]}[\sum_{i=1}^{4} b_{ii} - 2(b_{12} + b_{13} + b_{24} + b_{34} - b_{14} - b_{23})] \quad \text{(A10)}$$

where $\{b_{ij}\}$, $i, j = 1, \ldots, 4$ are given by a symmetric matrix $B = [b_{ij}]_{i,j=1}^{4} = (W_{[k]}^{-1})^T \Sigma_{[k]}^{*} W_{[k]}^{-1}$.

Equation (17b) follows directly from equation (A10). Equations (17c-d) follows directly from applying the independence between cases and controls to equations (17a-b).

**ACKNOWLEDGEMENT**

The author is grateful to the reviewer's comments which greatly improved the presentation of this paper.

**REFERENCES**

Acheson, E.D. and Gardner, M.J. (1979). Mesothelioma and exposure to mixtures of chrysotile and amphibole. *Arch. Env. Hlth.*, **34**, 240-242.

Agresti, A. (2002). *Categorical Data Analysis*. (2nd ed.). Wiley, New York. (Chapter 3).

Anderson, T.W. (2003). *An Introduction to Multivariate Statistical Analysis*. (3rd ed.) Wiley, New York. (Chapter 2).

Barron, B.A. (1977). The effects of misclassification on the estimation of relative risk. *Biometrics,* **33**, 414-418.

Bartlett, M.S. (1935). Contingency table interactions. *J. Roy. Statist. Soc.* (*Sup.*), **2**, 248-252.

Berry, G., Newhouse, M.L. and Turok, M. (1972). Combined effect of asbestos exposure and smoking on mortality from lung cancer in factory workers. *The Lancet*, (September 2), 476-479.

Bhapkar, V.P. and Koch, G.G. (1968). On the hypothesis of 'no interaction' in contingency tables. *Biometrics*, **24**, 567-594.

Birch, M.W. (1964). The detection of partial association, I: the 2 × 2 case. *J. Roy. Statist. Soc.,* **B26**, 313-324.

Boffetta, P., Winn, D.M., Ioannidis, J.P., Thomas, D.C., Little, J., Smith, G.D., Cogliano, V.J., Hecht, S.S., Seminara, D. Vineis, P. and Khoury, M.J. (2012). Reommendations and proposed guidelines for assessing the cumulative evidence on joint effects of gene and environments on cancer occurrence in humans. *Int. J. Epidemiol*., **41**, 1-19.

Brenner, H., Savitz, D.A. and Gefeller, O. (1993). The effects of joint misclassification of exposure and disease on epidemiologic measures of association. *J. Clinic. Epidemiol*., **46**, 1195-1202.

Chiacchierini, R.P. and Arnold, J.C. (1977). A two sample test for independence in 2 × 2 contingency tables with both margins subject to misclassification. *J. Amer. Statist. Assoc*., **72,** 170-174.

Cochran, W.G. (1968). Errors of measurement in statistics. *Technometrics*, **10,** 637-666.

Copeland, K.T., Checkoway, H., McMichael A.J. and Holbrook, R.H. (1977). Bias due to misclassification in the estimation of relative risk. *Amer. J. Epidemiol*., **105**, 488-495.

Cornfield, J. (1962). Join dependence of risk of coronary heart disease on serum cholesterol and systolic blood pressure: A discriminant function analysis. *Fed. Proc*., **21**, 58-61.

Darroch, J.N. and Borkent, M. (1994). Synergism, attributable risk and interaction for two binary exposure factors. *Biometrika,* **81**, 259-270.

Epstude, K. and Roese, N.J. (2008). The functional theory of counterfactual thinking. *Pers. Soc. Psychol. Rev.*, **12**, 168-192.

Flegal, K.M., Brownie, C. and Haas, J.D. (1986). The effect of exposure misclassification on estimates of relative risk. *Amer. J. Epidemiol.*, **123**, 736-751.

Fleiss, J., Levin, B. and Paik, M.C. (2003). *Statistical Methods for Rates and Proportions*. (3rd ed.), Wiley, New York. (Chapter 17).

Fung, K.Y. and Howe, G.R. (1984). Methodological issues in case-control studies III: The effect of joint misclassification of risk factors and confounding factors upon estimation and power. *Int. J. Epidemiol*., **13**, 366-370.

Garcia-Closas, M., Rothman, N. and Lubin, J. (1999). Misclassification in case-control studies of gene-environment interactions: Assessment of bias and sample size. *Cancer Epidemiol. Biomarkers Prev*., **8**, 1043-1050.

Gardner M.J. and Munford, A.G. (1980). The combined effect of two factors on disease in a case-control study. *J. Roy. Statist. Soc.,* **C29**, 276-281.

Keller, A.Z. and Terris, M. (1965). The association of alcohol and tobacco with cancer of the mouth and pharynx. *Am. J. Public Health*, **55**, 1578-1585.

Keys, A. and Kihlberg, J.K. (1963). Effects of misclassification on estimated relative prevalence of a characteristic Part II. Errors in two variables. *Amer. J. Public Health,* **53**, 1661-1665.

Kleinbaum, D.G., Kupper, L.L. and Morgenstern, H. (1982). *Epidemiologic Research: Principles and Quantitative Methods*. Wiley, New York. (Chapter 12).

Kristensen, P. (1992). Bias from non-differential but dependent misclassification of exposure and outcome. *Epidemiology,* **3**, 210-215.

Kuha, J., Skinner, C. and Palmgren, J. (2001). Misclassification error. In: *Encyclopedia of Biostatistics*, P. Armitage and T. Colton (eds), 2615-2621. Wiley, Chichester.

Kupper, L.L. and Hogan, M.D. (1978). Interaction in epidemiologic studies. *Am. J. Epidemiolo*., **108**, 447-454.

Lagakos, S.W. (1988). Effects of mismodelling and mismeasuring explanatory variables on tests of their association with a response variable. *Stat. Med*., **7**, 257-274.

Lee, P.N. (2001). Relation between exposure to asbestos and smoking jointly and the risk of lung cancer. *Occup. Environ. Med*., **58**, 145-153.

Lee, T-S. (2007). Correcting the estimation bias for joint misclassification errors from two binary variables, *Proceedings of the 4th Sino-International Symposium on Probability, Statistics, and Quantitative Management*, 129-152, Taipei, Taiwan.

Lee, T-S. (2009). Bias-adjusted exposure odds ratio for misclassified data. The Internet Journal of Epidemiology, **6(2)**, 1-19. Accessed from "http://www.ispub.com/journal/the-internet-journal-of-epidemiology/volume-6-number- 2/bias-adjusted-exposure-odds-ratio-for-misclassified-data-1.html".

MATLAB (2010). The Language of Technical Computing. www.mathworks.com.

Morrissey, M.J. and Spiegelman, D. (1999). Matrix methods for estimating odds ratios with misclassified exposure data: extensions and comparisons. *Biometrics*, **55**, 338-344.

Okamoto, K. and Horisawa, R. (2007). The joint effect of oxidative stress and antioxidants on the risk of an aneurysmal rupture subarachnoid hemorrhage: A case-control study in Japan. *Ann. Epidemiol.*, **17**, 359-363.

Ottman, R. (1996). Gene-environment interaction: Definitions and study designs. *Prev. Med.*, **25**, 764-770.

Pooley, F.D. (1976). An examination of the fibrous mineral content of asbestos lung tissue from Canadian chrysotile. *Environ. Res*., **12**, 281-298.

Rothman, K.J. (1974). Synergy and antagonism in cause-effect relationship. *Am. J. Epidemiol*., **99**, 385-388.

Rothman, K.J. (1976). The estimation of synergy or antagonism. *Am. J. Epidemiol*., **103**, 506-511.

Rothman, K.J. and Keller, A. (1972). The effect of joint exposure to alcohol and tobacco on risk of cancer on the mouth and pharynx. *J. Chron. Dis*., **25**, 711-716.

Rothman, K.J. and Greenland, S. (1998). *Modern Epidemiology*. (2nd ed.). Lippincott Williams and Wilkins, Philadelphia, PA. (Chapter 19).

Saracci, R. (1977). Asbestos and lung cancer: an analysis of the epidemiological evidence of the asbestos-smoking interaction. *Int. J. Cancer*, **20**, 323-331.

Tarafder, M.R., Carabin, H., McGarvey, S.T., Joseph, L., Balolong Jr., E. and Olveda, R. (2011). Assessing the impact of misclassification error on an epidemiological association between two helminthic infections. *PLoS Negl. Trop. Dis*., **5**, 1-7.

Thomas, D., Stram, D. and Dwyer, J. (1993). Exposure measurement error: Influence on exposure-disease relationships and methods of correction. *Ann. Rev. Public Health,* **14**, 69-93.

Tzonou, A., Kaldor, J., Smith, P.G., Day, N.E. and Trichopoulos, D. (1986). Misclassification in case-control studies with two dichotomous risk factors. *Revue d'Epidémiologie et de Santé Publique*, **34**, 10-17.

Vincent, R.G. and Marchetta, F. (1963). The relationship of the use of tobacco and alcohol to cancer of the oral cavity, pharynx or larynx. *Am. J. Surgery*, **106**, 501-505.

Walter, S.D. and Holford, T.R. (1978). Additive, multiplicative, and other models for disease risks. *Amer. J. Epidemiol.*, **108**, 341-346.

Walter, S.D. and Irwig, L.M. (1988). Estimation of test error rates, disease prevalence, and relative risk from misclassified data: A review. *J. Clinic. Epidemiol*., **41**, 923-937.