DigitalCommons
@WayneState

Journal of Modern Applied Statistical Methods

Volume 12 | Issue 2 Article 12

11-1-2013

Testing the Assumption of Non-differential Misclassification in Case-Control Studies

Tze-San Lee
Western Illinois University, Macomb, IL, leetzesan@gmail.com

Qin Hui
Emory University, Atlanta, GA, qin.hui@emory.edu

Testing the Assumption of Non-differential Misclassification in Case-Control Studies

Cover Page Footnote
Part of this paper is based on the result obtained in the Master degree thesis of the second author (QH) (Hui, 2011), which was written under the supervision of the first author (TL). The work in Figure 1 and Tables 3b, 4b and 5b is attributed to QH; the remaining work is attributed to TL. QH is grateful to TL who provided him with an invaluable guidance and constant encouragement while he was writing his thesis.

Journal of Modern Applied Statistical Methods
November 2013, Vol. 12, No. 2, 211-230.

ISSN 1538 − 9472

Testing the Assumption of Non-differential Misclassification in Case-Control Studies

Tze-San Lee
Western Illinois University
Macomb, IL

Qin Hui
Emory University
Atlanta, GA

One of the not yet solved issues regarding the misclassification in case-control studies is whether the misclassification rates are the same for both cases and controls. Currently, a common practice is to assume that the rates are the same, that is, the non-differential misclassification assumption. However, it has been suspected that this assumption may not be valid in practical applications. Unfortunately, no test is available so far to test the validity of the non-differential misclassification assumption. A method is presented to test the validity of non-differential misclassification assumption in case-control studies with 2 × 2 tables when validation data are not available. First, a theory of exposure operating characteristic curve is developed. Next, two non-parametric methods are presented to test the assumption of non-differential misclassification. Three real-data sets taken from practical applications are used as examples to illustrate the methods.

Keywords: Exposure operating characteristic (EOC) curve, non-differential misclassification, sensitivity, specificity, Youden's index

Introduction

One of the issues regarding the misclassification in case-control studies is whether the misclassification error rates are the same for both cases and controls (Walker & Irwig, 1988). Currently, a common practice is to assume that the rates are the same. This is the so-called "non-differential mis-classification (NDMC)" assumption. Many nice theoretical results are derived under this assumption. For example, in a case-control study with 2×2 contingency table, the adjusted odds ratio is always biased toward the value of the null hypothesis if the misclassification error rates are assumed to be non-differential (Bross, 1954; Goldberg, 1975). However, it is intuitively obvious that the assumption of NDMC might not be valid in many practical applications. Unfortunately, no test is available so far to test the validity of the NDMC assumption.

Dr. Lee is a mathematical statistician at the Centers for Disease Control and Prevention in Chamblee, GA. Email him at: tjl3@cdc.gov. Mr. Hui is an information analyst at the Rollins School of Public Health. Email him at qin.hui@emory.edu.

It is the purpose of this study to propose a test for assessing the validity of NDMC assumption in a case-control study with 2×2 contingency table. First, a theory of exposure operating characteristic curve is developed. Next, two methods are proposed to test the NDMC assumption. Three examples from practical applications are given to illustrate the proposed methods.

Methods

The curve of exposure operating characteristic

The idea of exposure operating characteristic (*EOC*) curve is parallel to that of receiver operating characteristic (*ROC*) curve in medical diagnostic test (Zhou, McClish & Obuchoowski, 2002). Suppose that the collected data for a case-control study is arranged to be given by Table 1. Assume that it is known that Table 1 is possibly misclassified; yet, the truly correct table is unknown. Here the counterfactual thinking comes into playing a crucial role in finding out what the possible true table is, that is, the true table is the counterfactual while the observed misclassified table is the factual (Epstude & Roese, 2008). It may thus be assumed that cell count in the observed table might be over- (or under-) misclassified by a certain number of subjects from the true table. The random variable E in Table 2 is assumed to be correctly classified on the subject's exposure condition, whereas E^* in Table 1 is its misclassified surrogate of E.

Table 1. The observed cell frequencies in a contingency table for a case-control study.

Classified exposure status	*Subject Group*	
	$Y = 1$ (Cases)	**$Y = 0$ (Controls)**
$E^* = 1$ (exposed)	n_{11}	n_{10}
$E^* = 0$ (unexposed)	n_{01}	n_{00}

Table 2. The [unobserved] true cell frequencies corresponding to Table 1.

Classified exposure status	*Subject Group*	
	$Y = 1$ (Cases)	**$Y = 0$ (Controls)**
$E = 1$ (exposed)	N_{11}	N_{10}
$E = 0$ (unexposed)	N_{01}	N_{00}

Let the number of misclassified subjects be given by

$$m_{(i)}^{(j)} = \text{the number of misclassified subjects between true and observed cell frequencies} = N_{ij} - n_{ij} \tag{1}$$

where $m_{(i)}^{(j)}$ (= ±1, ±2, ±3, …) is assumed and N_{ij} can be obtained as $N_{ij} = n_{ij} + m_{(i)}^{(j)}$. It will be clear how to choose the value of $m_{(i)}^{(j)}$ by applying the counterfactual thinking to the observed misclassified cell frequency as shown in the three examples of practical applications later in section 3.

The observed cell frequency (n_{ij}) is said to be under-misclassified if $m_{(i)}^{(j)} > 0$; otherwise it is called over-misclassified. Thus, the sensitivity (*Se*) and specificity (*Sp*) can be calculated for cases and controls as follows:

$$Se(m_{(1)}^{(j)}) = \Pr(E^* = 1 \mid E = 1; D = j) = 1 - \Pr(E^* = 0 \mid E = 1; D = j) = 1 - \frac{|m_{(1)}^{(j)}|}{N_{1j} + n_{1j}}$$

and (2)

$$Sp(m_{(0)}^{(j)}) = \Pr(E^* = 0 \mid E = 0; D = j) = 1 - \Pr(E^* = 1 \mid E = 0; D = j) = 1 - \frac{|m_{(0)}^{(j)}|}{N_{0j} + n_{0j}}$$

Note that not all $Se\left(m_{(1)}^{(j)}\right)$ and/or $Sp\left(m_{(0)}^{(j)}\right)$ are feasible. They have to satisfy the following three constraints which are imposed by the cell frequencies in Table 1 (Lee, 2009):

$$Se\left(m_{(1)}^{(j)}\right) + Sp\left(m_{(0)}^{(j)}\right) \neq 1 \tag{3a}$$

$$Se\left(m_{(1)}^{(j)}\right) > \hat{p}_j \tag{3b}$$

$$Sp\left(m_{(0)}^{(j)}\right) > \hat{q}_j \tag{3c}$$

where $\hat{p}_j = n_{1j} / \left(n_{1j} + n_{0j}\right)$ and $\hat{q}_j = 1 - \hat{p}_j$, j = 0, 1.

Varying the values of $m_{(i)}^{(j)}$, it is possible to obtain many feasible sensitivity and specificity pairs. A plot of all feasible pairs of points $\left(Se\left(m_{(1)}^{(j)}\right), 1 - Sp\left(m_{(0)}^{(j)}\right)\right)$ is said to be the *EOC* curve for cases or controls depending on j = 1 or 0. Incidentally, let the number of points on the *EOC* curves for controls and cases be given respectively by m_0 and m_1.

Testing the assumption of non-differential misclassification

In terms of the *EOC* curve, a test on the NDMC is equivalent to following pair of null and alternative hypotheses:

$$H_0 : EOC_1 = EOC_0 \text{ versus } H_1 : EOC_1 \neq EOC_0 \tag{4}$$

There are at least two ways to test equation 4. One way is to use a summary measure, the area under the curve (*AUC*). The measure of *AUC* has been widely used in testing whether the two *ROC* curves associated with the diseased and healthy populations are the same (Hanley & McNeil, 1982). The other way is to use the Kolmogorov-Smirnov test for the bivariate data of sensitivity and specificity pairs.

If a linear interpolation is used to connect all the discrete points on the *EOC* curve, the area under the curve is calculated by using numerical method, namely, the trapezoidal rule. For convenience, let $X(t)$ and $Y(t)$ denote respectively the x-axis (1 – Specificity) and y-axis (Sensitivity), where the variable t represents the misclassified number of subjects. The points $\left(x\left(m_{(0)k}^{(j)}\right), y\left(m_{(1)k}^{(j)}\right)\right)$ lying on the EOC_j curve are given respectively as follows: for j = 0, 1; k = 1, …, m_j.

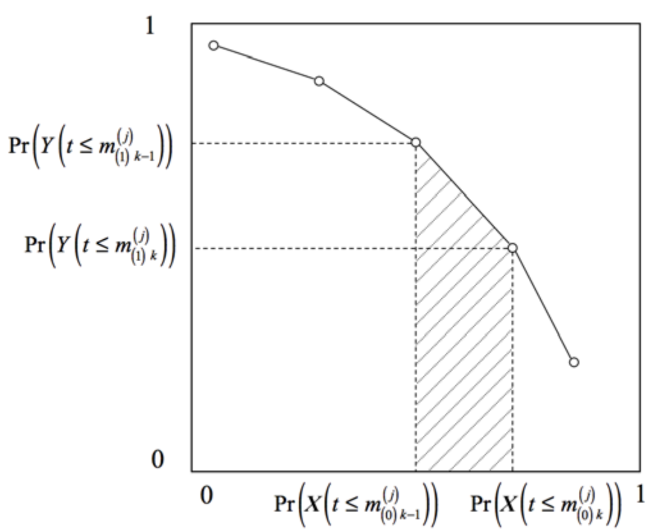

Figure 1. The ordinal dominance graph for the *EOC* curve

$$x\left(m_{(0)\,k}^{(j)}\right) = P\left(E^* = 1; t = m_{(0)\,k}^{(j)} \mid E = 0; D = j\right)$$

$$y\left(m_{(1)\,k}^{(j)}\right) = P\left(E^* = 1; t = m_{(1)\,k}^{(j)} \mid E = 1; D = j\right)$$

Thus, the EOC_j curve can be viewed as the ordinal dominance (OD) graph with each point on the EOC_j curve with $x\left(m_{(0)\,k}^{(j)}\right)$ and $y\left(m_{(1)\,k}^{(j)}\right)$ as its horizontal and vertical coordinates (Fig. 1). Thus, the area under the EOC_j curve (AUC) is calculated by using the trapezoidal rule as follows (Bamber, 1975):

$$\begin{aligned}
\theta \equiv AUC &= \sum_{k=1}^{m_j} \frac{1}{2}\left(y\left(m_{(1)\,k}^{(j)}\right) + y\left(m_{(1)\,k-1}^{(j)}\right)\right) \cdot \left(x\left(m_{(0)\,k}^{(j)}\right) - x\left(m_{(0)k-1}^{(i)}\right)\right) \\
&= \sum_{k=1}^{m_j} \frac{1}{2}\left(P\left(Y\left(t \le m_{(1)\,k}^{(j)}\right)\right) + P\left(Y\left(t \le m_{(1)\,k-1}^{(j)}\right)\right) \cdot P\left(X\left(m_{(0)\,k}^{(j)}\right)\right)\right) \\
&= \sum_{k=1}^{m_j} \left(P\left(Y\left(t \le m_{(1)\,k-1}^{(j)}\right)\right) + \frac{1}{2} P\left(Y\left(m_{(1)\,k}^{(j)}\right)\right) \cdot P\left(X\left(m_{(0)\,k}^{(j)}\right)\right)\right) \\
&= P\left(Y(t) < X(t)\right) + \frac{1}{2} P\left(Y(t) = X(t)\right)
\end{aligned} \tag{5}$$

To estimate equation 5, it can be shown that the AUC under the EOC is equivalent to the Wilcoxon-Mann-Whitney test (Pepe, 2003).

Two nonparametric methods for testing equation 4 are thereby summarized as follows:

Method A: The Wilcoxon-Mann-Whitney (WMW) test For each point lying on the EOC_i curve, define the Youden's index (YI) as follows (Zhou, McClish & Obuchowski, 2002):

$$YI\left(P^{(i)}\right) = Se\left(P^{(i)}\right) + Sp\left(P^{(i)}\right) - 1, \tag{6}$$

where $P^{(i)}$ is the point lying on the EOC_i curves, i = 0, 1.

Let $P_j^{(1)}$ and $Q_k^{(0)}$ be the points lying on the empirical EOC_i curves for i =1 (cases) and i = 0 (controls) respectively. Define

$$U_{jk} = 1 \text{ if } YI\left(P_j^{(1)}\right) > YI\left(Q_k^{(0)}\right); = 0 \text{ otherwise (assuming no ties).} \tag{7}$$

With equation 5 the null and alternative hypotheses of equation 4 are replaced by

$$H_0' : \theta = 0.5 \text{ versus } H_1' : \theta > 0.5 \tag{8}$$

An unbiased estimator for θ of equation 5 is given by Mee (1990)

$$\hat{\theta} = \frac{1}{m_1 \cdot m_0} \sum_{j=1}^{m_1} \sum_{k=1}^{m_0} U_{jk} \tag{9}$$

where U_{jk} are defined by equation 7 and its variance is given by

$$\operatorname{var}\left(\hat{\theta}\right) = \theta\left(1-\theta\right) / M, \tag{10}$$

where

$$M = m_0 m_1 / \left[\left(m_0 - 1\right)\delta_1 + \left(m_1 - 1\right)\delta_2 + 1\right]$$

and for $\ell = 1, 2$,

$$\delta_\ell = (\theta_\ell - \theta^2) / (\theta(1-\theta)),$$

$$\theta_1 = \Pr(U_{ij} U_{kj} = 1), i \neq k,$$

$$\theta_2 = \Pr(U_{ij} U_{ik} = 1), j \neq k.$$

Note that the estimators for θ_1, θ_2, δ_ℓ, and M are

$$\hat{\theta}_1 = \sum_{i=1}^{m_0} \sum_{j=1}^{m_1} \sum_{k \neq i}^{m_0} U_{ij} U_{kj} / \left[m_0 m_1 \left(m_0 - 1\right)\right]$$

$$\hat{\theta}_2 = \sum_{i=1}^{m_0}\sum_{j=1}^{m_1}\sum_{k\neq j}^{m_1} U_{ij}U_{ik} / \left[m_0 m_1 \left(m_1 - 1\right)\right]$$

$$\tilde{\theta}^2 = \left[m_0 m_1 \hat{\theta}^2 - \left(m_0 - 1\right)\hat{\theta}_1 - \left(m_1 - 1\right)\hat{\theta}_2 - \hat{\theta}\right] / \left[\left(m_0 - 1\right)\left(m_1 - 1\right)\right], \quad (11)$$

$$\tilde{\delta}_\ell = \left(\hat{\theta}_\ell - \tilde{\theta}^2\right) / \left(\hat{\theta} - \tilde{\theta}^2\right)$$

$$\tilde{M} = m_0 m_1 / \left[\left(m_0 - 1\right)\tilde{\delta}_1 + \left(m_1 - 1\right)\tilde{\delta}_2 + 1\right] = \left(\hat{\theta} - \tilde{\theta}^2\right) / \left(\hat{\theta}^2 - \tilde{\theta}^2\right)$$

Consequently, an estimator of $\operatorname{var}\left(\hat{\theta}\right)$ is given by

$$\widehat{\operatorname{var}}\left(\hat{\theta}\right) = \hat{\theta}\left(1 - \hat{\theta}\right) / \tilde{M} \quad (12)$$

Hence, a standard normal z_θ-statistic for testing equation 8 is given by

$$z_\theta = \frac{\hat{\theta} - 0.5}{\sqrt{\widehat{\operatorname{var}}\left(\hat{\theta}\right)}} \quad (13)$$

Method B: The Kolmogorov-Smirnov test Let S_{m_1} and T_{m_0} be the sample cumulative distribution function of Youden's index (equation 6) associated with the number of points lying on the *EOC* curves for cases and controls respectively, where S_{m_1} and T_{m_0} are defined respectively as

$$\begin{aligned} S_{m_1}(t) &= 0, && t < YI\left(P^{(1)}\right)_{(1)}, \\ &= k / m_1 && YI\left(P^{(1)}\right)_{(k)} \le t \le YI\left(P^{(1)}\right)_{(k+1)}, \\ &= 1, && t \ge YI\left(P^{(1)}\right)_{(m1)}; \end{aligned} \quad (14)$$

$$\begin{aligned} T_{m_0}(t) &= 0, && t < YI\left(Q^{(0)}\right)_{(1)}, \\ &= k / m_0 && YI\left(Q^{(0)}\right)_{(k)} \le t \le YI\left(Q^{(0)}\right)_{(k+1)}, \\ &= 1, && t \ge YI\left(Q^{(0)}\right)_{(m0)}, \end{aligned} \quad (15)$$

where $YI(P^{(1)})_{(i)}$, i = 1,2, …, m_1, and $YI(Q^{(0)})_{(j)}$, j = 1, 2, …, m_0 are the order statistics of Youden's index (equation 6) associated with points lying on the *EOC* curves for cases and controls respectively.

The Kolmogorov-Smirnov test is based on the statistic $K_{m_0 m_1}$ defined as

$$K_{m_0 m_1} = \sup_t |S_{m_1}(t) - T_{m_0}(t)|, \tag{16}$$

A decision rule for testing equation 4 is given as follows: reject the null hypothesis of equation 4 if the observed number ($K_{m_0 m_1}$) (equation 16) is larger than the two-sided critical value K_α, where α is the probability of type I error (Conover, 1971).

Examples

Three examples are used to illustrate how to employ the two methods mentioned in the previous section to test the assumption of non-differential misclassification. The problem now is to calculate the value of sensitivity and specificity when the validation sample data are not available. Here the counterfactual thinking comes into playing the critical role to overcome this barrier (Epstude & Roese, 2008), that is, if only the true (correctly classified) table is known, it is then possible to calculate the value of sensitivity and specificity pair from the observed [misclassified] table by regarding the true table which serves as the "gold standard." Evidently, the potential true table, even though unknown, can be figured out from the observed table as shown below in each of the following three examples. Because it is unknown which potential outcome table is the genuine true table, it is necessary to consider all possible outcome tables figuring out from the observed table as the true table. This leads to a plot of the *EOC* curve separately for the over-/under-misclassification situation in all three examples.

Because the critical values of $K_{0.05}$ are not available for all the following three examples, it was calculated using the large sample approximation $1.36\sqrt{(m_0 + m_1)/(m_0 m_1)}$, which provided in the last row of Table 17 in (Conover, 1971).

Example 1 The data in Table 3a are taken from a study of deaths caused by landslides that occurred in the State of Chuuk, Federated States of Micronesia, in which a case-control design was used to identify the risk factors (Sanchez et al., 2009). A case was defined to be a person who died as a result of landslides.

Proxies were identified in the surviving villagers to provide information for the decedents, or persons in the control group who were too young to answer questions. Because proxies were used to obtain information on the questions asked in the survey, misclassification was likely to occur. For an illustration, one table was taken from their study regarding whether a person saw natural warning signs (Table 3a). In this example, Exposure = 1 if a person did not see natural warning signs; 0 otherwise.

Assume that the observed table (Table 3a) is misclassified, the potential true table for the over-misclassification situation may be determined by identifying all possible positive integers less than the smallest observed frequency in the (0, 1) cell, n_{01} = 2. It turns out there is only one integer which is less than 2. Hence the only potential true (counterfactual) table is given by N_{11} = 38, N_{10} = 26, N_{01} =1, and N_{00} = 26. By using equation 2, $Se_1 = 1 - 1/(38 + 37) = 0.987$ and $Sp_1 = 1 - 1/(1 + 2) = 0.667$; $Se_0 = 1 - 1/(26 + 27) = 0.981$ and $Sp_0 = 1 - 1/(26 + 25) = 0.98$. Hence, the *EOC* curves for cases and controls have just one point $(Se_1, 1 - Sp_1)$ = (0.987, 0.333) and $(Se_0, 1 - Sp_0)$ = (0.981, 0.02) as shown in Table 3b. Although there were 24 true (counterfactual) tables for the under-misclassification situation, only three and seventeen sensitivity and specificity pairs for cases and controls were proved respectively to be feasible, namely, they satisfy all the three constraints of Eqs. 3a-3c. All feasible (1 – *Sp*, *Se*) pairs are exhibited as boldface figures in Table 3b. A plot of the *EOC* curves for cases and controls in Example 1 is given in Fig. 2.

Because the results of both methods are not significant (Table 3c), the null hypothesis of equation 4 is not rejected at the significance level of 0.05.

Table 3a. Survey data: whether or not a person saw natural warning signs for cases and controls

	Cases	Controls	*Total*
No	37	27	***64***
Yes	2	25	***27***
Total	***39***	***52***	***91***

Table 3b. True (counterfactual) table and the corresponding feasible sensitivity and specificity

Cases				Controls			
Over-misclassification							
N_{11}	***N_{01}***	***Se_1***	***Sp_1***	***N_{10}***	***N_{00}***	***Se_0***	***Sp_0***
38	1	0.9867	0.6667	26	26	0.9811	0.9804
Under-misclassification							
N_{11}	***N_{01}***	***Se_1***	***Sp_1***	***N_{10}***	***N_{00}***	***Se_0***	***Sp_0***
13	26	0.5200	0.1429	51	1	0.6923	0.0769
14	25	0.5490	0.1481	50	2	0.7013	0.1481
15	24	0.5769	0.1538	49	3	0.7105	0.2143
16	23	0.6038	0.1600	48	4	0.7200	0.2759
18	21	0.6545	0.1739	46	6	0.7397	0.3871
19	20	0.6786	0.1818	45	7	0.7500	0.4375
20	19	0.7018	0.1905	44	8	**0.7606**	**0.4848**
21	18	0.7241	0.2000	43	9	**0.7714**	**0.5294**
22	17	0.7458	0.2105	42	10	**0.7826**	**0.5714**
23	16	0.7667	0.2222	41	11	**0.7941**	**0.6111**
24	15	0.7869	0.2353	40	12	**0.8060**	**0.6486**
25	14	0.8065	0.2500	39	13	**0.8182**	**0.6842**
26	13	0.8254	0.2667	38	14	**0.8308**	**0.7179**
27	12	0.8438	0.2857	37	15	**0.8438**	**0.7500**
28	11	0.8615	0.3077	36	16	**0.8571**	**0.7805**
29	10	0.8788	0.3333	35	17	**0.8710**	**0.8095**
30	9	0.8955	0.3636	34	18	**0.8852**	**0.8372**
31	8	0.9118	0.4000	33	19	**0.9000**	**0.8636**
32	7	0.9275	0.4444	32	20	**0.9153**	**0.8889**
33	6	0.9429	0.5000	31	21	**0.9310**	**0.9130**
34	5	**0.9577**	**0.5714**	30	22	**0.9474**	**0.9362**
35	4	**0.9722**	**0.6667**	29	23	**0.9643**	**0.9583**
36	3	**0.9863**	**0.8000**	28	24	**0.9818**	**0.9796**

Table 3c. Result of applying Methods A and B to Example 1

Method A: Wilcoxon-Mann-Whitney test				**Method B: Kolmogorov-Smirnov test**	
m_0	18	$\tilde{\delta}_1$	-0.05	m_0	18
m_1	4	$\tilde{\delta}_2$	0.17	m_1	4
$\hat{\theta}_1$	0.22	$\tilde{M}$	19	$K_{m_1 m_0}$	0.33
$\hat{\theta}_2$	0.28	$\hat{\theta}$	0.5	$K_{0.05}$	0.75
$\tilde{\theta}^2$	0.24	$\sqrt{\hat{\text{var}}\left(\hat{\theta}\right)}$	0.11		
		$z_{\hat{\theta}}$	0		

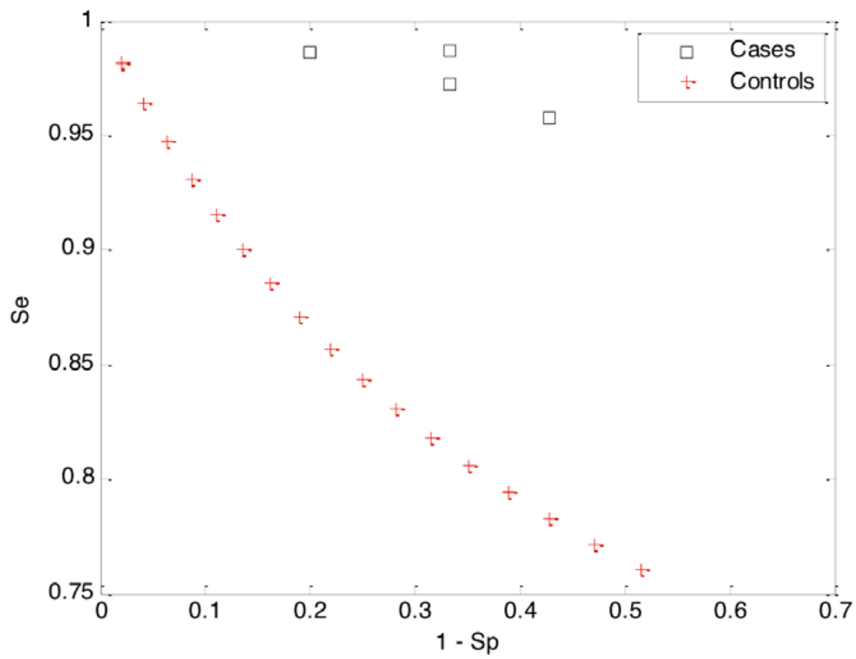

Figure 2. *EOC* curves for cases and controls in Example 1

Example 2 The data in Table 4a are taken from table 7.2 in Schlesselman's book (1982). This data set is in fact a subset of the data from a case-control study of the relation between estrogen use and endometrial cancer in women (Antunes et al., 1979). The use of estrogen is regarded as an exposure risk factor. The rates of exposure for cases and controls are given respectively by $\hat{p}_1$ = 0.3 (= 55/183) and $\hat{p}_0$ = 0.1 (= 19/183). Assume that the exposure data are misclassified for both cases and controls and there is interest in knowing whether their misclassification rates are the same.

By designating the frequency in cell (1, 0) of Table 4a as a free parameter, there were 18 potential true (counterfactual) tables for the over-misclassification scenario. After checking for the feasibility constraints (Eqs. 3b-3c), all 18 pairs of sensitivity and specificity were feasible for cases, while only 17 pairs were feasible for controls. For the under-misclassification scenario, there were 64 potential true (counterfactual) tables. Yet 45 pairs of sensitivity and specificity for cases were feasible, while 30 pairs were feasible for controls. Again, only the top and bottom five pairs are listed in Table 4b. A plot of their *EOC* curves is given in Fig. 3.

Because the results of both methods are not significant (Table 4c), the null hypothesis of equation 4 is not rejected at the significance level of 0.05. By the way, the reason that $\hat{\theta} = 0.42$ < 0.5 is because equation 7 is defined in terms of controls rather than cases, that is, $U_{jk} = 1$ if $YI\left(Q_j^{(0)}\right) > YI\left(P_k^{(1)}\right)$.

Table 4a. Use of oral conjugated estrogen (*OCE*) for endometrial cancer

	Cases	Controls	*Total*
User	55	19	***74***
Nonuser	128	164	***292***
Total	***183***	***183***	***366***

Table 4b. True (counterfactual) table and the corresponding feasible sensitivity and specificity

Cases				**Controls**			
Over-misclassification							
N_{11}	N_{01}	Se_1	Sp_1	N_{10}	N_{00}	Se_0	Sp_0
54	129	0.9908	0.9961	20	163	0.9744	0.9970

Table 4b Continued

Cases				Controls			
Over-misclassification							
N_{11}	**N_{01}**	**Se_1**	**Sp_1**	**N_{10}**	**N_{00}**	**Se_0**	**Sp_0**
53	130	0.9815	0.9922	21	162	0.9500	0.9939
52	131	0.9720	0.9884	22	161	0.9268	0.9908
51	132	0.9622	0.9846	23	160	0.9048	0.9877
50	133	0.9524	0.9808	24	159	0.8837	0.9845
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
29	154	0.6905	0.9078	45	138	0.5938	0.9139
28	155	0.6747	0.9046	46	137	0.5846	0.9103
27	156	0.6585	0.9014	47	136	0.5758	0.9067
26	157	0.6420	0.8982	48	135	0.5672	0.9030
25	158	0.6250	0.8951	49	134	0.5588	0.8993
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
14	169	0.4058	0.8620	60	123	-*	-*
13	170	0.3823	0.8591	61	122	-*	-*
12	171	0.3582	0.8562	62	121	-*	-*
11	172	0.3333	0.8533	63	120	-*	-*
10	173	0.3077	0.8505	64	119	-*	-*
Under-misclassification							
N_{11}	**N_{01}**	**Se_1**	**Sp_1**	**N_{10}**	**N_{00}**	**Se_0**	**Sp_0**
73	110	0.8594	0.9244	1	182	-*	-*
72	111	0.8661	0.9289	2	181	0.1905	0.9507
71	112	0.8730	0.9333	3	180	0.2727	0.9535
70	113	0.8800	0.9378	4	179	0.3478	0.9563
69	114	0.8871	0.9421	5	178	0.4167	0.9591
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
60	123	0.9565	0.9800	14	169	0.8485	0.9850
59	124	0.9649	0.9841	15	168	0.8824	0.9880
58	125	0.9735	0.9881	16	167	0.9143	0.9909
57	126	0.9821	0.9921	17	166	0.9444	0.9939
56	127	0.9910	0.9961	18	165	0.9730	0.9970

*The values of (Se, Sp) are infeasible.

Table 4c. Result of applying the two methods to Example 2

Method A: Wilcoxon-Mann-Whitney test				**Method B:** Kolmogorov-Smirnov test	
m_0	63	$\tilde{\delta}_1$	0.46	m_0	63
m_1	47	$\tilde{\delta}_2$	0.23	m_1	47
$\hat{\theta}_1$	0.23	$\tilde{M}$	82.1	$K_{m_1 m_0}$	0.21
$\hat{\theta}_2$	0.29	$\hat{\theta}$	0.42	$K_{0.05}$	0.26
$\tilde{\theta}^2$	0.17	$\sqrt{\hat{\text{var}}\left(\hat{\theta}\right)}$	0.05		
		$z_{\hat{\theta}}$	-1.44		

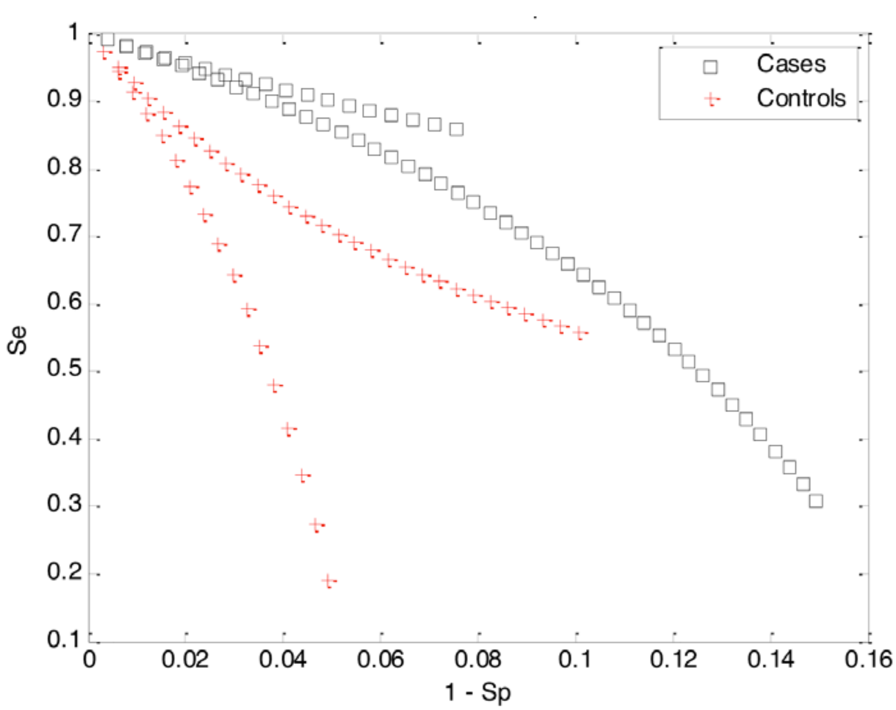

Figure 3. *EOC* curves for cases and controls in Example 2

Example 3 The data of Table 5a are taken from a case control study of sudden infant death syndrome (SIDS) (Greenland, 1988). Among those women who were interviewed, it was asked if she had used the antibiotic medicine during pregnancy. The rate of using the antibiotic medicine for cases and controls were given respectively by $\hat{p}_1$ = 0.22 (= 122/442) and $\hat{p}_0$ = 0.17 (= 101/580). Assume that the interview data are misclassified for both cases and controls and there is interest in knowing whether their misclassification rates are the same.

To do so, it is necessary to obtain their *EOC* curves. To construct the potential true (counterfactual) table, the observed cell frequency n_{10} = 101 were chosen as a reference. For the over-misclassification scenario, the possible values of N_{10} were determined to be integers running from 100 down to 1 (Column 5, Table 5b). After the value of N_{10} was determined, all other cell frequencies were uniquely determined because the column/row totals have to be fixed as the same as that of the observed table. There were 100 potential true (counterfactual) tables. After checking the feasibility constraints imposed by Eqs. 3b-c, all 100 (*Se*, *Sp*) pairs were feasible for cases, but only 91 (*Se*, *Sp*) pairs were feasible for controls. To save space, only the top and bottom five pairs are listed (Table 5b). Similarly, for the under-misclassification scenario, the possible values of N_{10} were determined to be integers running from 102 up to 222. There were 121 potential true (counterfactual) tables. Although all 121 true (counterfactual) tables produced feasible pairs of (*Se, Sp*) for controls, only 107 (*Se, Sp*) pairs were feasible for cases. Again, only the top and bottom five pairs are listed (Table 5b). A plot of their *EOC* curves for cases and controls is given respectively in Fig. 4.

Because none of the results obtained from both methods are significant (table 5c), the null hypothesis of equation 4 is not rejected at the significance level of 0.05.

Table 5a. Data of SIDS study of the exposure variable of interview response

	Cases	Controls	*Total*
Use	122	101	***223***
No use	442	479	***921***
Total	***564***	***580***	***1144***

Table 5b. True (counterfactual) table and the corresponding feasible sensitivity and specificity

Cases				Controls			
			Over-misclassification				
N_{11}	**N_{01}**	**Se_1**	**Sp_1**	**N_{10}**	**N_{00}**	**Se_0**	**Sp_0**
123	441	0.9960	0.9989	100	480	0.9950	0.9990
124	440	0.9919	0.9977	99	481	0.9900	0.9979
125	439	0.9879	0.9966	98	482	0.9849	0.9969
126	438	0.9839	0.9955	97	483	0.9798	0.9958
127	447	0.9799	0.9943	96	484	0.9746	0.9948
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
218	346	0.7176	0.8782	14	566	0.2435	0.9167
219	345	0.7155	0.8768	13	567	0.2281	0.9159
220	344	0.7135	0.8754	12	568	0.2124	0.9150
221	343	0.7114	0.8739	11	569	0.1964	0.9141
222	342	0.7093	0.8724	10	570	0.1802	0.9133
			Under-misclassification				
N_{11}	**N_{01}**	**Se_1**	**Sp_1**	**N_{10}**	**N_{00}**	**Se_0**	**Sp_0**
121	443	0.9959	0.9989	102	478	0.9951	0.9990
120	444	0.9917	0.9977	103	477	0.9902	0.9979
119	445	0.9876	0.9966	104	476	0.9854	0.9969
118	446	0.9833	0.9955	105	475	0.9806	0.9958
117	447	0.9791	0.9944	106	474	0.9758	0.9948
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
19	545	0.2695	0.8956	204	376	0.6623	0.8795
18	546	0.2571	0.8947	205	375	0.6601	0.8782
17	547	0.2446	0.8938	206	374	0.6580	0.8769
16	548	0.2319	0.8929	207	373	0.6558	0.8756
15	549	0.2190	0.8920	208	372	0.6537	0.8743
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
5	559	-*	-*	218	362	0.6332	0.8609
4	560	-*	-*	219	361	0.6313	0.8595
3	561	-*	-*	220	360	0.6293	0.8582
2	562	-*	-*	221	359	0.6273	0.8568
1	563	-*	-*	222	358	0.6254	0.8554

*The values of (Se, Sp) are infeasible.

Table 5c. Result of applying the two methods to Example 3

Method A: Wilcoxon-Mann-Whitney test				Method B: Kolmogorov-Smirnov test	
m_0	212	$\tilde{\delta}_1$	0.33	m_0	212
m_1	207	$\tilde{\delta}_2$	0.37	m_1	207
$\hat{\theta}_1$	0.38	$\tilde{M}$	296.9	$K_{m_1 m_0}$	0.05
$\hat{\theta}_2$	0.39	$\hat{\theta}$	0.54	$K_{0.05}$	0.13
$\tilde{\theta}^2$	0.29	$\sqrt{\hat{\text{var}}\left(\hat{\theta}\right)}$	0.03		
		$z_{\hat{\theta}}$	1.52		

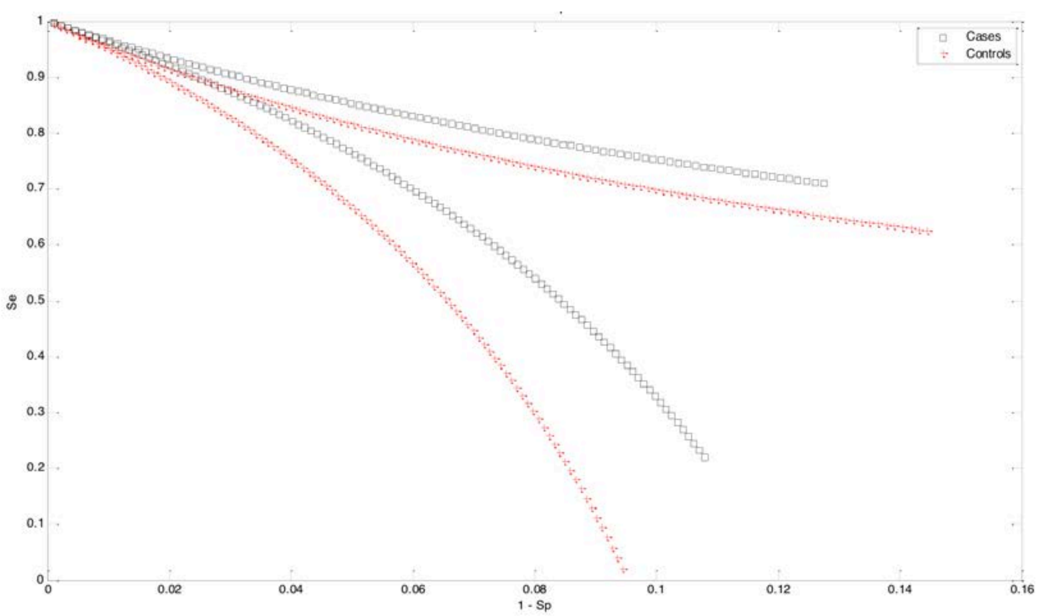

Figure 4. *EOC* curves for cases and controls in Example 3

Discussion

Some comments are worthy to be mentioned below:

1) The *EOC* curve is intrinsically different from that of the *ROC* curve. The *ROC* curve is interested in judging the accuracy of a diagnostic test on the individual's disease status, while the *EOC* curve is concerned with the correct classification of the subject's exposure condition.

2) Unlike the *ROC* curve in which the entire curve is a single continuous curve, the *EOC* curve is comprised of two distinct pieces: one piece of the curve corresponds to the over-misclassification scenario, while the other piece of the curve to the under-misclassification. Further, the *ROC* curve is strictly increasing, whereas the *EOC* curve is monotonically decreasing.

3) It seems that equation 3a is redundant when both equations 3b &3c are satisfied. But, the expression of $Se\left(m_1^{(j)}\right)+Sp\left(m_0^{(j)}\right)-1$ is the determinant of the misclassification matrix for the 2 × 2 contingency table. In fact, equation 3a is the first condition required for the existence of the bias-adjusted proportion estimator (Lee, 2009). Incidentally, the non-singularity of the misclassification matrix is always the first condition required to be satisfied for the existence of the bias-adjusted estimator in other applications too (Lee, 2010, 2011).

4) Method B is preferred to Method A because it is possible that two *EOC* curves are different, but they have the same area.

Conclusion

In this paper a theory of the exposure operating characteristic curve is developed to test the assumption of non-differential misclassification in case-control studies. In terms of the Youden's index two nonparametric methods, the Wilcoxon-Mann-Whitney and Kolmogorov-Smirnov test, are proposed to test whether the two exposure operating characteristic curves are the same for cases and controls. Three real-data examples were used to illustrate the proposed two methods.

Apparently, the idea of the exposure operating characteristic curve for testing the assumption of non-differential misclassification for the 2 × 2 contingency tables presented can be extended to the 2 × *K* or *K* × *K* matched-pair case-control studies, where $K \geq 3$. This topic will be pursued later in another paper.

Acknowledgements

Part of this paper is based on the result obtained in the Master degree thesis of the second author (QH) (Hui, 2011), which was written under the supervision of the first author (TL). The work in Figure 1 and Tables 3b, 4b and 5b is attributed to QH; the remaining work is attributed to TL. QH is grateful to TL who provided him with an invaluable guidance and constant encouragement while he was writing his thesis.

References

Antunes, C. M. F., Stolley, P. D., Rosenshein, N. B., Davies, J. L., Tonascia, J. A., Brown, C., Burnett, L., Rutledge, A., Pokempner, M., & Garcia, R. (1979). Endometrial cancer and estrogen use: Report of a large case-control study. *New England Journal of Medicine*, *300*, 9-13.

Bamber, E. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, *12*, 387-415.

Bross, I. (1954). Misclassification in 2 × 2 tables. *Biometrics*, *10*, 478-486.

Conover, W. J. (1971). *Practical Nonparametric Statistics*. New York: John Wiley & Sons.

Epstude, K., & Roese, N. J. (2008). The functional theory of counterfactual thinking. *Personnel Social and Psychological Review*, *12*, 168-192.

Goldberg, J. D. (1975). The effects of misclassification on the bias in the difference between two proportions and the relative odds in the fourfold table. *Journal of American Statistical Association*, *70*, 561-567.

Greenland, S. (1988). Variance estimation for epidemiologic effect estimates under misclassification. *Statistics in Medicine*, *7*, 745-757.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (*ROC*) curve. *Radiology*, *143*, 29-36.

Hui, Q. (2011). *Testing an Assumption of Nondifferential Misclassification in Case-Control Studies*. Master Degree Thesis, Georgia State University, Atlanta, GA.

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

Lee, T-S. (2010). Misclassified ordinal data in case-control studies. *Journal of Probability and Statistical Science*, *8*, 215-226.

Lee, T-S. (2011). Matched-pair studies with misclassified ordinal data. *Journal of Modern Applied Statistical Methods*, *10*, 67-76.

Mee, R. W. (1990). Confidence intervals for probabilities and tolerance regions based on ageneralization of the Mann-Whitney statistic. *Journal of American Statistical Association*, *85*, 793-800.

Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. New York: Oxford University Press.

Sanchez, C., Lee, T-S., Young, S., Batts, D., Benjamin, J., & Malilay, J. (2009). Risk factors for mortality during 2002 landslides in the State of Chuuk, Federated Staes of Micronesia. *J Disasters*, *33*, 705-720.

Schlesselman, J. J. (1982). *Case-Control Studies: Design, Conduct, Analysis*. Oxford, England: Oxford University Press.

Walker, S. D., & Irwig, L. M. (1988). Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review. *Journal of Clinical Epidemiology*, *41*, 923-937.

Zhou, X-H., McClish, D. K., & Obuchowski, N. A. (2002). *Statistical Methods in Diagnostic Medicine*. New York: Wiley.