Georgia State University
Digital Archive @ GSU

Mathematics Theses | Department of Mathematics and Statistics

8-1-2011

Testing an Assumption of Non-Differential Misclassification in Case-Control Studies

Qin Hui
Georgia State University, qhui1@student.gsu.edu

Follow this and additional works at: http://digitalarchive.gsu.edu/math_theses

Recommended Citation
Hui, Qin, "Testing an Assumption of Non-Differential Misclassification in Case-Control Studies" (2011). *Mathematics Theses.* Paper 103.

This Thesis is brought to you for free and open access by the Department of Mathematics and Statistics at Digital Archive @ GSU. It has been accepted for inclusion in Mathematics Theses by an authorized administrator of Digital Archive @ GSU. For more information, please contact digitalarchive@gsu.edu.

TESTING AN ASSUMPTION OF NON-DIFFERENTIAL MISCLASSIFICATION IN CASE-CONTROL STUDIES

by

QIN HUI

Under the Direction of Tzesan Lee and Gengsheng Qin

ABSTRACT

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. Currently, a common practice is to assume that the rates are the same ("non-differential" assumption). However, it is suspicious that this assumption is valid in many case-control studies. Unfortunately, no test is available so far to test the validity of the assumption of non-differential misclassification when the validation data are not available. We propose the first such method to test the validity of non-differential assumption in a case-control study with 2 × 2 contingency table. First, the Exposure Operating Characteristic curve is defined. Next, two non-parametric methods are applied to test the assumption of non-differential misclassification. Three examples from practical applications are used to illustrate the methods and a comparison is made.

KEY WORDS: Case-Control studies, Exposure operating characteristic, Non-differential misclassification, Sensitivity, Specificity

TESTING AN ASSUMPTION OF NON-DIFFERENTIAL MISCLASSIFICATION IN CASE-CONTROL STUDIES

by

QIN HUI

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Science

In the College of Arts and Sciences

Georgia State University

2011

Copyright by

Qin Hui

2011

TESTING AN ASSUMPTION OF NON-DIFFERENTIAL MISCLASSIFICATION IN CASE-CONTROL STUDIES

by

QIN HUI

Committee Chair: Dr. Gengsheng (Jeff) Qin

Co-Chair: Dr. Tzesan Lee

Member: Dr. Xu Zhang

Electronic Version Approved:

Office of Graduate Studies

College of Arts and Sciences

Georgia State University

August 2011

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to the following persons for their suggestion and assistance in the preparation of this thesis. Dr. Tzesan Lee of CDC served as my thesis advisor, gave me invaluable guidance and constant encouragement. Without his direction and supervision, this thesis would not have been possibly written. Dr. Gengsheng (Jeff) Qin of GSU, as my joint advisor, spent many hours in shaping the direction of my study, offering me concrete suggestions on various topics.

Last but not least, I would like to thank my family and my friends who gave me constant support and encouragement through the whole process.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS ……iv

LIST OF TABLES ……..….. vii

LIST OF FIGURES …….……... ix

CHAPTER 1 INTRODUCTION 1

1.1 Misclassification ……...…….…. 1

1.2 Methods to Correct the Misclassification ……...……..……... 2

1.3 Non-differential Assumption …….……..….. 2

1.4 Validity of Non-differential Assumption in Case-Control Study …….…. 4

1.5 This Work ……..……..……... 5

CHAPTER 2 METHODS AND PROCEDURES…….…. 7

2.1 The Curve of Exposure Operating Characteristic (EOC)…….…. 7

2.2 Testing Assumption of Non-differential Misclassification ……..……... 10

2.3 The Original Data …….……..…... 13

CHAPTER 3 REAL APPLICATIONS …….…...…..…... 15

3.1 Sensitivity and Specificity …….…...…..… 15

3.2 The Empirical EOC Curves ……..…… 33

3.3 Hypothesis Testing ……..... 36

3.3.1 Method A …….……..... 36

3.3.2 Method B …….……..….. 38

CHAPTER 4 DISCUSSIONS …….…… 45

REFERENCES …….…….…….. 47

APPENDIX …….…….…….….. 49

LIST OF TABLES

Table 1 The observed cell counts for a case-control study …….. 7

Table 2 The unobserved true cell counts for the case-control study of Table 1 …… 8

Table 3 The survey data whether a person saw natural warning signs …….…….. 13

Table 4 The data of SIDS study of the exposure variable of interview ……. 14

Table 5 Use of oral conjugated estrogen (OCE) for endometrial cancer ……...…. 14

Table 6 Feasible 1-Sp and Se for under-misclassification of Example 1, Case ……. 17

Table 7 Feasible 1-Sp and Se for under-misclassification of Example 1, Control …… 17

Table 8 Feasible 1-Sp and Se for over-misclassification of Example 1, Case ……... 18

Table 9 Feasible 1-Sp and Se for over-misclassification of Example 1, Control …….. 18

Table 10 Feasible 1-Sp and Se for under-misclassification of Example 2, Case ……..... 19

Table 11 Feasible 1-Sp and Se for under-misclassification of Example 2, Control ..…… 22

Table 12 Feasible 1-Sp and Se for over-misclassification of Example 2, Case ……..... 25

Table 13 Feasible 1-Sp and Se for over-misclassification of Example 2, Control …… 27

Table 14 Feasible 1-Sp and Se for under-misclassification of Example 3, Case ……... 29

Table 15 Feasible 1-Sp and Se for under-misclassification of Example 3, Control …….. 30

Table 16 Feasible 1-Sp and Se for over-misclassification of Example 3, Case ……... 30

Table 17 Feasible 1-Sp and Se for over-misclassification of Example 3, Control …… 31

Table 18 The result of Mann-Whiney Statistics for Example 1 ……..……... 37

Table 19 The result of Mann-Whiney Statistics for Example 2 …….…… 37

Table 20 The result of Mann-Whiney Statistics for Example 3 …….……..….. 38

Table 21 The result of Graphical Method for Example 1 …….……..…… 40

Table 22 The result of Graphical Method for Example 2 …….……..…… 42

Table 23 The result of Graphical Method for Example 3 …….……..…….. 43

LIST OF FIGURES

Figure 1 EOC curves of Example 1, under-misclassification ……..... 33

Figure 2 EOC curves of Example 1, over-misclassification …….…..... 34

Figure 3 EOC curves of Example 2, under-misclassification ……..…… 34

Figure 4 EOC curves of Example 2, over-misclassification …… 35

Figure 5 EOC curves of Example 3, under-misclassification ……..…… 35

Figure 6 EOC curves of Example 3, over-misclassification …… 36

Figure 7 Fitting results for Example 1, under-misclassification …… 39

Figure 8 Fitting results for Example 1, over-misclassification …… 40

Figure 9 Fitting results for Example 2, under-misclassification ……..... 41

Figure 10 Fitting results for Example 2, over-misclassification ……. 41

Figure 11 Fitting results for Example 3, under-misclassification …… 42

Figure 12 Fitting results for Example 3, over-misclassification ……. 43

CHAPTER 1 INTRODUCTION

1.1 Misclassification

Misclassification in epidemiology means a subject's (exposure or disease) status is mistakenly classified. For example, a patient exposed to a risk factor may be recorded as nonexpose by mistake. Almost all epidemiologic studies suffer, more or less, from misclassification (Walker and Irwig, 1988). In general, this problem may be caused by the following reasons (Espeland and Hui, 1987):

(1) Misreporting by study subjects. For observational studies using self reported variables, the subjects may not remember correctly, or intentionally hide the true status.

(2) The use of less than perfect measurement devices/methods. Any diagnostic instrument or method has certain level of false positive and false negative rates. This means not all true patients will be diagnosed as having the disease, or vice versa.

(3) Random error. Even if the diagnostic devices/methods are perfect, the testing results may still be wrong due to random errors.

In observational epidemiologic applications, the exposure misclassification occurs more often when surrogate respondents are used in the survey interview to classify the subject's exposure status. Misclassification occurs if the information obtained from the proxy does not reflect the patient's exposure status correctly (Nelson et al., 1990). For example, in a case-control study of possible etiologic factors for Alzheimer's disease, information was obtained only from close family members, because of the patient's mental impairment (Rocca et al., 1986). The authors assessed the reliability of proxy respondents to provide interview data for the specific items of the case-control study of Alzheimer's disease conducted in Italy. For all questions of the interview, responses of 52 non-demented subjects were compared to responses of their close

relatives. Although the authors found it encouraging for the use of next-of-kin respondents in case-control studies of Alzheimer's disease or other neurologic conditions for which the subject cannot provide historical information, 21–27% of the pairs the next-of-kin was unable to answer questions about general anesthesia, antacid drug use, and age of mother and father at index birth. Though the surrogate respondent was able to answer 45 of 57 tested items, the overall agreement was about 80%. Questions about use of hard liquor and behavior pattern yielded agreement of 71–75%, while those about number of jobs, and number of cigarettes per day yielded only 62–63% agreement. For those who provided information about antacid drug use, agreement was even poorer. In other words, the misclassification could be more than 40% for certain questions.

1.2 Methods to Correct Misclassification

The problem of misclassification in epidemiology was first studied by Bross (1954), and later thoroughly investigated by Barron (1977), Chen (1989), Chu et al., (2006), among others. Various methods have been developed to evaluate and account for the effect of misclassification on study validity. When misclassification error rates, or sensitivity and specificity, are known from, for example, literature, experts, or can be directly estimated from an internal validation subsample of the study subjects, mathematical formula or computer programs are available to correct the misclassification and reveal the true prevalence or exposure counts.

1.3 Non-differential Assumption

To use the above methods to analyze misclassified data, one must assume that the exposure misclassification error rates are identical for all the subgroups of the study population. In other words, the exposure misclassification error rates are independent of the disease status. This is the

so-called “non-differential” assumption (Greenland, 1988). On the other hand, if the misclassification rates are different between subgroups, it is called “differential”. Many theoretical results are derived under the “non-differential” assumption. For example, in a case-control study with 2 × 2 contingency table, the adjusted odds ratio is always biased toward the value stated in the null hypothesis if the misclassification error rates are assumed to be non-differential.

The “non-differential” assumption may be a good approximation in many studies, but is questionable in case-control studies. For example, the cases (patients) may be more likely to recall exposure than the controls (non-patients), resulting in a higher sensitivity or lower specificity for cases. As a consequence, data analysis results based on non-differential assumption may be totally wrong as shown in the following example (Greenland, 1988; Drews et al., 1987).

In a case-control study of sudden infant death syndrome (SIDS), it was found that, 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 (see Table 4; this example will be used to test our proposed method later in this thesis). The uncorrected data based on interview yield an odds ratio of 1.31 with 95% confidence interval (0.98, 1.76), indicating a positive association between antibiotic use and SIDS. Medical records were also employed in this study as validation data, which is considered much more reliable than the interview data, though is not perfect. The validation study yields an estimated sensitivity of 0.6304 for case mothers, and only 0.5676 for control mothers, and an estimated specificity of 0.8667 for case mothers but 0.9333 for non-case mothers, clearly showing that the non-differential assumption is not valid in this example. Using non-differential assumption yields a corrected odds ration of 1.73 with 95%

confidence interval (0.92, 3.24), in sharp contrast to a corrected odds ratio of only 0.73, with 95% confidence interval (0.25, 2.15) using differential error rates. Therefore, after correction for differential misclassification, the data show no evidence regarding the association between antibiotic use and SIDS, in contrary to the result yielded from uncorrected data or corrected assuming non-differential misclassification. This example demonstrated clearly the importance of testing the validity of the non-differential assumption in case-control studies.

1.4 Validity of Non-differential Assumption in Case-Control Study

Even though under some special situation a “gold standard”, e.g. medical records in above example, may be used to determine the true exposure status and exam the validity of the non-differential assumption, in most case-control studies, however, it is not practical to do so, either because the validation data cannot be obtained, or the cost is too high. This can be best illustrated by the following example (Sanchez et al., 2009). This example will be used as Example 1 later in this thesis.

In a case-control study of deaths caused by landslides that occurred in the State of Chuuk, Federated States of Micronesia, neighbors and relatives were identified as proxies for 40 victims who were died as results of the landslides, to provide information on whether or not they saw natural signs of landslides, while 53 survivors were employed as control group. It is obvious that no validation data could be obtained for the case group in this example. Even if misclassification error rate can be estimated for the controls by some means, it is questionable if this error rate can be applied to the case group before the non-differential assumption is proved to be valid for this study.

Thus it is desirable to develop a method which can exam the validity of the non-differential assumption, using only the data from the main study (i.e. without the validation data). If the test result is positive, then validation data, if available, obtained from a subgroup, say, control, can be applied to the whole population. If the test result is negative, meaning the misclassification is differential, the validation data must be obtained for both case and control, respectively, or the analysis of the data may be flawed. Unfortunately, to the best of our knowledge, no such method is available up to date.

1.5 This Work

Recently a method has been proposed to calculate the true error rates when a dichotomous exposure variable is misclassified in a case-control study, based only on the data available in the main study (Lee, 2009). This method obtained the truly classified contingency table by assuming hypothetically that it is simply a table which is misclassified from the observed one by 1, 2, 3… subjects. The sensitivity (and/or specificity) was thus calculated from the observed and the true table. A bias-adjusted exposure odds ratio with its asymptotic variance was then presented to account for the misclassification bias. Following that work, in this study we propose the first test method which is able to exam the validity of non-differential assumption in a case-control study with 2 × 2 contingency table.

This thesis is arranged as follows: The Exposure Operating Characteristic (EOC) curve is defined in Chapter 2. Two non-parametric methods are applied to the EOC curves to test the non-differential assumption. In Chapter 3, three examples from practical applications are used to

illustrate the proposed methods and a comparison is made among these methods. Finally some conclusion and discussion are presented in Chapter 4.

CHAPTER 2 METHODS AND PROCEDURES

2.1 The Exposure Operating Characteristic Curve

The idea of the Exposure Operating Characteristic (EOC) curve is parallel but intrinsically different from that of the ROC curve in medical diagnostic test (Zhou et al., 2002). Suppose that the collected data for a case-control study is given in Table 1. All we know is that Table 1 is possibly misclassified. We don't know what the true (or correctly classified) table is, but we can assume that a difference between the true table (Table 2) and the observed table (Table 1) was that cells in the observed table were over- or under-misclassified by a certain number of subjects from the true table. Let this misclassified number of subjects be $m_{(i)j}$ given by equation 1.

Table 1 The observed cell counts for a case-control study.

Classified exposure status	Group of subjects	
	D = 1 (Cases)	D = 0 (Controls)
$E^* = 1$ (exposed)	n_{11}	n_{10}
$E^* = 0$ (unexposed)	n_{01}	n_{00}
Total observed sample size	n_{+1}	n_{+0}

Table 2 The unobserved true cell counts for the case-control study of Table 1.

Classified exposure status	Group of subjects	
	D = 1 (Cases)	D = 0 (Controls)
E = 1 (exposed)	N_{11}	N_{10}
E = 0 (unexposed)	N_{01}	N_{00}
Total true sample size	N_{+1}	N_{+0}

Note that $N_{+j} = n_{+j}$, j = 0, 1. The three variables E, E*, and D are defined as follows:

E = 1 if a subject is truly exposed, 0 otherwise;

E* = 1 if a subject is classified as exposed, 0 otherwise;

D = 1 if a subject belongs to the case group, 0 otherwise.

Let

$$m_{(i)j} = \text{the misclassified number between observed and true cell counts} = n_{ij} - N_{ij}. \qquad (1)$$

Note that if $m_{(i)j} \neq 0$, the observed table (Table 1) is said to be misclassified; otherwise, Table 1 is said to be correctly classified, that is, not misclassified. When the observed table is supposed to be misclassified, we need to calculate the probability of misclassification. Yet, we still do not know what the true table is. But that is all right. We can borrow some idea from the theory of counterfactual in causality analysis, that is, we can view the observed table as a factual one. Thus, the true table (unobserved) is regarded as a counterfactual one (or potential outcome). As a result, many true tables become available depending on the different values of $m_{(i)j}$. At this point,

we note that $m_{(i)j}$ is in fact at our disposition and N_{ij} can then be obtained as $N_{ij} = n_{ij} - m_{(i)j}$. Furthermore, a misclassification is called under-misclassified if $m_{(1)j} < 0$ (or $m_{(0)j} > 0$), meaning the observed number of exposed subjects is less than the true number of exposed subjects; otherwise it is called over-misclassified. Now, the sensitivity (*Se*) and specificity (*Sp*), of the classification procedure for cases and controls can be defined as follows:

$$Se = P(E^* = 1 \mid E = 1) = 1 - P(E^* = 0 \mid E = 1),$$

and (2)

$$Sp = P(E^* = 0 \mid E = 0) = 1 - P(E^* = 1 \mid E = 0).$$

For different values of $m_{(i)j}$ (= ±1, ±2, ±3, ….) we can estimate Se and Sp of equation 2 as follows:

$$Se_j(m_{(1)j}) = 1 - \frac{|m_{(1)j}|}{N_{1j} + n_{1j}},$$

and (3)

$$Sp_j(m_{(0)j}) = 1 - \frac{|m_{(0)j}|}{N_{0j} + n_{0j}},$$

where j = 0 (control), 1 (case).

Equation 3 is said to be feasible under the following constraints (Lee, 2009):

$$Se_j(m_{(1)j}) > \hat{p}_j \qquad (4)$$

$$Sp_j(m_{(0)j}) > \hat{q}_j \tag{5}$$

and

$$Se_j(m_{(1)j}) + Sp_j(m_{(0)j}) > 1, \tag{6}$$

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

The plot of all feasible pairs of (1-$Sp_j(m_{(0)j})$, $Se_j(m_{(1)j})$) for different X = $m_{(1)j}$ or Y = $m_{(0)j}$ is then called the *Exposure Operating Characteristic* (EOC) curve for either cases or controls, respectively.

2.2 Testing Assumption of Non-differential Misclassification

Let $EOC_j(X, Y)$, j = 0, 1, be a function of X and Y representing, respectively, the EOC curve for cases and controls. In terms of the function $EOC_j(X, Y)$, j = 0, 1, a null hypothesis for the assumption of non-differential misclassification is formulated as follows:

$$H_0 : EOC_0(X,Y) = EOC_1(X,Y). \tag{7a}$$

An alternative hypothesis is then given by

$$H_1 : EOC_0(X,Y) \neq EOC_1(X,Y). \tag{7b}$$

Two methods, Methods A and B, for testing H_0 versus H_1, are proposed as follows. Method A employs the area under the EOC curve, while Method B is a graphical method.

Method A: The area under the EOC curve

Let P_j and Q_k be the points lying on the EOC curve for cases and controls respectively. Define

$S(P_j, Q_k) = 1$ if P_j lies above Q_k; 0.5 if $P_j = Q_k$; 0 if P_j lies below Q_k. (8)

The area under the EOC curve is then given by

$$\theta_{MW} = \frac{1}{m_{K_0} \cdot m_{K_1}} \sum_{j=1}^{m_{K_1}} \sum_{k=1}^{m_{K_0}} S(P_j, Q_k), \tag{9}$$

where $S(P_j, Q_k)$ is given by equation 8. θ_{MW} is a statistic of Mann-Whitney type.

Assume that θ_{MW} follows a normal distribution. A standardized normal z_{MW}-statistic for testing H_0 versus H_1 is given by (Sprent and Smeeton, 2007)

$$z_{MW} = \frac{\theta_{MW} - \frac{1}{2}}{\sqrt{(m_{K_0} + m_{K_1} + 1)/(12 m_{K_0} m_{K_1})}}. \tag{10}$$

Method B: A Graphical method (Tsutakawa and Hewett, 1977)

Given the two empirical EOC curves, EOC_1 and EOC_0, for cases and controls, we plot another curve which is a least square fitting to the combined data from the EOC_1 and EOC_0. Then we count the number of points (V) from the EOC_1 which lies above the obtained least square fitted curve. By conditioning on that the sum (m_{K+}) of m_{K_1} and m_{K_0} is fixed, the conditional distribution of V follows a hypergeometric distribution, that is,

$$P(V = v) = \frac{\binom{m_{K_1}}{v} \cdot \binom{m_{K_0}}{m - v}}{\binom{m_{K+}}{m}}, \tag{11}$$

where m is the total number of points on both EOC_1 and EOC_0 which lie above the least square fitted curve, and $m_{K+} = m_{K_0} + m_{K_1}$. A decision rule for testing equations 7a-7b is given as follows: reject the null hypothesis of equation 7a if the observed number (v) is large.

2.3 Original Data

In this work we use three examples to demonstrate the application of our method for testing the assumption of non-differential misclassification. Example 1 was from a case-control study of mortality caused by landslides that occurred in the State of Chuuk, Federated States of Micronesia, during Tropical Storm Chata'an July 2-4, 2002 (Sanchez et al., 2009). 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 the decedents, or persons in the control group who are too young to answer questions. 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. For an illustrative purpose we only take one table from their study regarding whether a person saw natural warning signs (Table 3). Please note that in this example, "exposure" is defined as if some one did NOT see any natural warning sign ("No Signs" in Table 3), and "no exposure" is defined as if some one did see natural warning signs ("Signs" in Table 3).

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

Natural Warnings	Case	Control	Total
No Signs	37	27	64
Signs	2	25	27
Marginal Total	39	52	91

The second set of data used here, as shown in Table 4, is taken from a case-control study on sudden infant death syndrome (SIDS) (Greenland 1988, Drews et al., 1987). A case is defined as

a mother whose child suffered SIDS, and a control is defined as a mother whose child survived. Exposure is defined as if the mother used antibiotics during pregnancy ("Use" in Table 4), and no exposure is defined as if a mother did not use antibiotics during pregnancy ("No Use" in Table 4). The study found that 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 (see Table 4).

Table 4 The data of SIDS study of the exposure variable of interview response

Interview Response	Cases	Controls	Total
Use	122	101	223
No Use	442	479	921
Marginal Total	564	580	1144

Table 5 shows the third data set, which is taken from Table 7.2 in the book "Case-Control Studies: Design, Conduct, Analysis", by Schlesselman (1982). This is a subset of data from a large scale case-control study of endometrial cancer and estrogen use (Antunes et al., 1979). A case is defined as an endometrial cancer patient, while a control is a matched hospital patient without endometrial cancer. Exposure is defined as if the patient is an oral conjugated estrogen (OCE) user ("User" is Table 5), and no exposure is defined as if the patient is not an OCE user ("Nonuser" in Table 5).

Table 5 Use of oral conjugated estrogen (OCE) for endometrial cancer

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

CHAPTER 3 REAL APPLICATIONS

3.1 Sensitivity and Specificity

We take Example 2 to show how to calculate feasible sensitivity and specificity. From Table 4, we have

$$\hat{p}_1 = 122/564 = 0.216\,, \quad \hat{q}_1 = 1 - \hat{p}_1 = 0.784\,.$$

$$\hat{p}_0 = 101/580 = 0.174\,, \quad \hat{q}_0 = 1 - \hat{p}_0 = 0.826\,.$$

Since $N_{11} + N_{01} = 564$ and $N_{10} + N_{00} = 580$, we observe that N_{01} must be over-misclassified if N_{11} is under-misclassified with respect to the observed cell value. Calculations of sensitivity and specificity (Tables 6-17) will stop once either equation 3 or 4 is violated. Note that the absolute values for $m_{(i)j}$ which can also be read from Tables 6-17 are given, respectively, by $0 < m_{(1)1}$ (= $m_{(0)1}$) < 159 (= 281 – 122) for part (a), $0 < m_{(1)1}$ (= $m_{(0)1}$) < 106 (= 121 – 15) for part (b), $0 < m_{(1)0}$ (= $m_{(0)0}$) < 139 (= 240 – 101) for part (c), and $0 < m_{(1)0}$ (= $m_{(0)0}$) < 91 (= 100 – 9).

Incidentally, we're also able by using equations 2 to 5, to find the feasible range of m_{ij} for under- and over-misclassification for cases and controls with respect to "use" as follows:

For cases,

a. Under-misclassification for "use" and over-misclassification for "no use":

   $Se_1 = 1 - (N_{11} - 122)/(N_{11} + 122) > 0.216$ implies $122 < N_{11} < 1008$;

   But, because $N_{11} < 564$, we therefore have $122 < N_{11} < 564$;

   $Sp_1 = 1 - (442 - N_{01})/(N_{01} + 442) > 0.784$ implies $285 < N_{01} < 442$;

b. Over-misclassification for "use" and under-misclassification for "no use":

   $Se_1 = 1 - (122 - N_{11})/(N_{11} + 122) > 0.216$ implies $15 < N_{11} < 122$;

$Sp_1 = 1 – (N_{01} - 442)/(N_{01} + 442) > 0.784$ implies $442 < N_{01} < 686$;

But since $N_{01} < 564$, we have $442 < N_{01} < 564$.

For controls,

c. Under-misclassification for "use" and over-misclassification for "no use":

$Se_0 = 1 - (N_{10} – 101)/(N_{10} + 101) > 0.174$ implies $101 < N_{10} < 1060$;

But, since $N_{10} < 580$, we have $101 < N_{10} < 580$;

$Sp_0 = 1 – (479 - N_{00})/(N_{00} + 479) > 0.826$ implies $340 < N_{00} < 479$;

d. Over-misclassification for "use" and under-misclassification for "no use":

$Se_0 = 1 - (101 - N_{10})/(N_{10} + 101) > 0.174$ implies $10 < N_{10} < 101$;

$Sp_0 = 1 – (N_{00} - 479)/(N_{00} + 479) > 0.826$ implies $479 < N_{00} < 681$;

Since the sample size for cases and controls is assumed to be fixed, the feasibility for sensitivity and specificity has to be considered simultaneously. Hence the range of feasibility for the $m_{(i)j}$-values has to be chosen to satisfy the feasibility for sensitivity and specificity simultaneously for either under- or over-misclassification. Take the under-misclassification for "use" for cases as an example to illustrate how to determine the final feasible range for $m_{(1)1}$ and $m_{(0)1}$. Since the range for $m_{(1)1}$ between the observed value of 122 and its maximum feasible bound (564) which is 442 (= 564 – 122) which is greater than 157 (= 442 - 285), the range for $m_{(0)1}$, we have to choose the smaller value between m_{11} and m_{01}. Consequently, the feasible range for both $m_{(1)1}$ and m_{01} in part (a) is given by $0 < m_{(1)1}$ (= $m_{(0)1}$) < 157, while the feasible range for both $m_{(1)1}$ and $m_{(0)1}$ in part (b) is $0 < m_{(1)1}$ (= $m_{(0)1}$) < 107. Similarly, the feasible range in part (c) for controls is $0 < m_{(1)0}$ (= $m_{(0)0}$) < 139, while the feasible range in part (d) for controls is $0 < m_{(1)0}$ (= $m_{(0)0}$) < 91. Although there is a minor discrepancy for cases (parts (a)-(b)) between

the above calculations and that of the tables, this is because we only used three decimal digits for sensitivity and specificity calculations.

Table 6 and 7 list all the calculated feasible pairs of sensitivity (Se) and specificity (Sp) for cases and controls, respectively, for Example 1, where only under-misclassification for exposed subjects ("No Signs") is assumed to exist. Under-misclassification means the observed number of exposed subjects is less than the true number (as indicated by a negative $m_{(1)j}$, j=1,0). For convenience, values of (1-Sp) instead of Sp are listed.

Table 6 Feasible 1-Sp and Se for under-misclassification of Example 1, Case.

$m_{(1)1}$	No Signs	Signs	1-Sp	Se
-1	38	1	0.333333	0.986667

Table 7 Feasible 1-Sp and Se for under-misclassification of Example 1, Control.

$m_{(1)0}$	No Signs	Signs	1-Sp	Se
-17	44	8	0.515152	0.760563
-16	43	9	0.470588	0.771429
-15	42	10	0.428571	0.782609
-14	41	11	0.388889	0.794118
-13	40	12	0.351351	0.80597
-12	39	13	0.315789	0.818182
-11	38	14	0.282051	0.830769
-10	37	15	0.25	0.84375
-9	36	16	0.219512	0.857143
-8	35	17	0.190476	0.870968
-7	34	18	0.162791	0.885246
-6	33	19	0.136364	0.9
-5	32	20	0.111111	0.915254
-4	31	21	0.086957	0.931034

-3	30	22	0.06383	0.947368
-2	29	23	0.041667	0.964286
-1	28	24	0.020408	0.981818

Table 8 and 9 list all the calculated feasible pairs of sensitivity (Se) and specificity (Sp) for cases and controls, respectively, for Example 1, where only over-misclassification for exposed subjects ("No Signs") is assumed to exist. Over-misclassification means the observed number of exposed subjects is more than the true number (as indicated by a positive $m_{(1)j}$, j=1,0). For convenience, again values of (1-Sp) instead of Sp are listed.

Table 8 Feasible 1-Sp and Se for over-misclassification of Example 1, Case.

$m_{(1)1}$	No Signs	Signs	1-Sp	Se
+1	36	3	0.2	0.986301
+2	35	4	0.333333	0.972222
+3	34	5	0.428571	0.957746

Table 9 Feasible 1-Sp and Se for over-misclassification of Example 1, Control.

$m_{(1)0}$	No Signs	Signs	1-Sp	Se
+1	26	26	0.019608	0.981132
+2	25	27	0.038462	0.961538
+3	24	28	0.056604	0.941176
+4	23	29	0.074074	0.92
+5	22	30	0.090909	0.897959
+6	21	31	0.107143	0.875
+7	20	32	0.122807	0.851064

+8	19	33	0.137931	0.826087
+9	18	34	0.152542	0.8
+10	17	35	0.166667	0.772727
+11	16	36	0.180328	0.744186
+12	15	37	0.193548	0.714286
+13	14	38	0.206349	0.682927
+14	13	39	0.21875	0.65
+15	12	40	0.230769	0.615385
+16	11	41	0.242424	0.578947
+17	10	42	0.253731	0.540541

Table 10 and 11 list all the calculated feasible pairs of sensitivity (Se) and specificity (Sp) for cases and controls, respectively, for Example 2, where only under-misclassification for exposed subjects is assumed to exist. Under-misclassification means the observed number of exposed subjects is less than the true number (as indicated by a negative $m_{(1)j}$, j=1,0). For convenience, values of (1-Sp) instead of Sp are listed.

Table 10 Feasible 1-Sp and Se for under-misclassification of Example 2, Case.

$m_{(1)1}$	Use	No use	1-Sp	Se		$m_{(1)1}$	Use	No use	1-Sp	Se
-159	281	283	0.21931	0.605459		-80	202	362	0.099502	0.753086
-158	280	284	0.217631	0.606965		-79	201	363	0.098137	0.755418
-157	279	285	0.215956	0.608479		-78	200	364	0.096774	0.757764
-156	278	286	0.214286	0.61		-77	199	365	0.095415	0.760125
-155	277	287	0.21262	0.611529		-76	198	366	0.094059	0.7625
-154	276	288	0.210959	0.613065		-75	197	367	0.092707	0.76489
-153	275	289	0.209302	0.61461		-74	196	368	0.091358	0.767296
-152	274	290	0.20765	0.616162		-73	195	369	0.090012	0.769716

-151	273	291	0.206003	0.617722		-72	194	370	0.08867	0.772152
-150	272	292	0.20436	0.619289		-71	193	371	0.087331	0.774603
-149	271	293	0.202721	0.620865		-70	192	372	0.085995	0.77707
-148	270	294	0.201087	0.622449		-69	191	373	0.084663	0.779553
-147	269	295	0.199457	0.624041		-68	190	374	0.083333	0.782051
-146	268	296	0.197832	0.625641		-67	189	375	0.082007	0.784566
-145	267	297	0.196211	0.627249		-66	188	376	0.080685	0.787097
-144	266	298	0.194595	0.628866		-65	187	377	0.079365	0.789644
-143	265	299	0.192982	0.630491		-64	186	378	0.078049	0.792208
-142	264	300	0.191375	0.632124		-63	185	379	0.076736	0.794788
-141	263	301	0.189771	0.633766		-62	184	380	0.075426	0.797386
-140	262	302	0.188172	0.635417		-61	183	381	0.074119	0.8
-139	261	303	0.186577	0.637076		-60	182	382	0.072816	0.802632
-138	260	304	0.184987	0.638743		-59	181	383	0.071515	0.805281
-137	259	305	0.1834	0.64042		-58	180	384	0.070218	0.807947
-136	258	306	0.181818	0.642105		-57	179	385	0.068924	0.810631
-135	257	307	0.18024	0.643799		-56	178	386	0.067633	0.813333
-134	256	308	0.178667	0.645503		-55	177	387	0.066345	0.816054
-133	255	309	0.177097	0.647215		-54	176	388	0.06506	0.818792
-132	254	310	0.175532	0.648936		-53	175	389	0.063779	0.821549
-131	253	311	0.173971	0.650667		-52	174	390	0.0625	0.824324
-130	252	312	0.172414	0.652406		-51	173	391	0.061224	0.827119
-129	251	313	0.170861	0.654155		-50	172	392	0.059952	0.829932
-128	250	314	0.169312	0.655914		-49	171	393	0.058683	0.832765
-127	249	315	0.167768	0.657682		-48	170	394	0.057416	0.835616
-126	248	316	0.166227	0.659459		-47	169	395	0.056153	0.838488
-125	247	317	0.16469	0.661247		-46	168	396	0.054893	0.841379
-124	246	318	0.163158	0.663043		-45	167	397	0.053635	0.844291
-123	245	319	0.161629	0.66485		-44	166	398	0.052381	0.847222
-122	244	320	0.160105	0.666667		-43	165	399	0.05113	0.850174
-121	243	321	0.158585	0.668493		-42	164	400	0.049881	0.853147

-120	242	322	0.157068	0.67033		-41	163	401	0.048636	0.85614
-119	241	323	0.155556	0.672176		-40	162	402	0.047393	0.859155
-118	240	324	0.154047	0.674033		-39	161	403	0.046154	0.862191
-117	239	325	0.152542	0.6759		-38	160	404	0.044917	0.865248
-116	238	326	0.151042	0.677778		-37	159	405	0.043684	0.868327
-115	237	327	0.149545	0.679666		-36	158	406	0.042453	0.871429
-114	236	328	0.148052	0.681564		-35	157	407	0.041225	0.874552
-113	235	329	0.146563	0.683473		-34	156	408	0.04	0.877698
-112	234	330	0.145078	0.685393		-33	155	409	0.038778	0.880866
-111	233	331	0.143596	0.687324		-32	154	410	0.037559	0.884058
-110	232	332	0.142119	0.689266		-31	153	411	0.036342	0.887273
-109	231	333	0.140645	0.691218		-30	152	412	0.035129	0.890511
-108	230	334	0.139175	0.693182		-29	151	413	0.033918	0.893773
-107	229	335	0.137709	0.695157		-28	150	414	0.03271	0.897059
-106	228	336	0.136247	0.697143		-27	149	415	0.031505	0.900369
-105	227	337	0.134788	0.69914		-26	148	416	0.030303	0.903704
-104	226	338	0.133333	0.701149		-25	147	417	0.029104	0.907063
-103	225	339	0.131882	0.70317		-24	146	418	0.027907	0.910448
-102	224	340	0.130435	0.705202		-23	145	419	0.026713	0.913858
-101	223	341	0.128991	0.707246		-22	144	420	0.025522	0.917293
-100	222	342	0.127551	0.709302		-21	143	421	0.024334	0.920755
-99	221	343	0.126115	0.71137		-20	142	422	0.023148	0.924242
-98	220	344	0.124682	0.71345		-19	141	423	0.021965	0.927757
-97	219	345	0.123253	0.715543		-18	140	424	0.020785	0.931298
-96	218	346	0.121827	0.717647		-17	139	425	0.019608	0.934866
-95	217	347	0.120406	0.719764		-16	138	426	0.018433	0.938462
-94	216	348	0.118987	0.721893		-15	137	427	0.017261	0.942085
-93	215	349	0.117573	0.724036		-14	136	428	0.016092	0.945736
-92	214	350	0.116162	0.72619		-13	135	429	0.014925	0.949416
-91	213	351	0.114754	0.728358		-12	134	430	0.013761	0.953125
-90	212	352	0.11335	0.730539		-11	133	431	0.0126	0.956863

-89	211	353	0.11195	0.732733		-10	132	432	0.011442	0.96063
-88	210	354	0.110553	0.73494		-9	131	433	0.010286	0.964427
-87	209	355	0.109159	0.73716		-8	130	434	0.009132	0.968254
-86	208	356	0.107769	0.739394		-7	129	435	0.007982	0.972112
-85	207	357	0.106383	0.741641		-6	128	436	0.006834	0.976
-84	206	358	0.105	0.743902		-5	127	437	0.005688	0.97992
-83	205	359	0.10362	0.746177		-4	126	438	0.004545	0.983871
-82	204	360	0.102244	0.748466		-3	125	439	0.003405	0.987854
-81	203	361	0.100872	0.750769		-2	124	440	0.002268	0.99187
						-1	123	441	0.001133	0.995918

Table 11 Feasible 1-Sp and Se for under-misclassification of Example 2, Control.

$m_{(1)0}$	Use	No use	1-Sp	Se		$m_{(1)0}$	Use	No use	1-Sp	Se
-139	240	340	0.169719	0.592375		-70	171	409	0.078829	0.742647
-138	239	341	0.168293	0.594118		-69	170	410	0.077615	0.745387
-137	238	342	0.16687	0.59587		-68	169	411	0.076404	0.748148
-136	237	343	0.16545	0.597633		-67	168	412	0.075196	0.750929
-135	236	344	0.164034	0.599407		-66	167	413	0.073991	0.753731
-134	235	345	0.162621	0.60119		-65	166	414	0.072788	0.756554
-133	234	346	0.161212	0.602985		-64	165	415	0.071588	0.759398
-132	233	347	0.159806	0.60479		-63	164	416	0.070391	0.762264
-131	232	348	0.158404	0.606607		-62	163	417	0.069196	0.765152
-130	231	349	0.157005	0.608434		-61	162	418	0.068004	0.768061
-129	230	350	0.155609	0.610272		-60	161	419	0.066815	0.770992
-128	229	351	0.154217	0.612121		-59	160	420	0.065628	0.773946
-127	228	352	0.152828	0.613982		-58	159	421	0.064444	0.776923
-126	227	353	0.151442	0.615854		-57	158	422	0.063263	0.779923
-125	226	354	0.15006	0.617737		-56	157	423	0.062084	0.782946
-124	225	355	0.148681	0.619632		-55	156	424	0.060908	0.785992

-123	224	356	0.147305	0.621538		-54	155	425	0.059735	0.789063
-122	223	357	0.145933	0.623457		-53	154	426	0.058564	0.792157
-121	222	358	0.144564	0.625387		-52	153	427	0.057395	0.795276
-120	221	359	0.143198	0.627329		-51	152	428	0.056229	0.798419
-119	220	360	0.141836	0.629283		-50	151	429	0.055066	0.801587
-118	219	361	0.140476	0.63125		-49	150	430	0.053905	0.804781
-117	218	362	0.13912	0.633229		-48	149	431	0.052747	0.808
-116	217	363	0.137767	0.63522		-47	148	432	0.051592	0.811245
-115	216	364	0.136418	0.637224		-46	147	433	0.050439	0.814516
-114	215	365	0.135071	0.639241		-45	146	434	0.049288	0.817814
-113	214	366	0.133728	0.64127		-44	145	435	0.04814	0.821138
-112	213	367	0.132388	0.643312		-43	144	436	0.046995	0.82449
-111	212	368	0.131051	0.645367		-42	143	437	0.045852	0.827869
-110	211	369	0.129717	0.647436		-41	142	438	0.044711	0.831276
-109	210	370	0.128386	0.649518		-40	141	439	0.043573	0.834711
-108	209	371	0.127059	0.651613		-39	140	440	0.042437	0.838174
-107	208	372	0.125734	0.653722		-38	139	441	0.041304	0.841667
-106	207	373	0.124413	0.655844		-37	138	442	0.040174	0.845188
-105	206	374	0.123095	0.65798		-36	137	443	0.039046	0.848739
-104	205	375	0.12178	0.660131		-35	136	444	0.03792	0.852321
-103	204	376	0.120468	0.662295		-34	135	445	0.036797	0.855932
-102	203	377	0.119159	0.664474		-33	134	446	0.035676	0.859574
-101	202	378	0.117853	0.666667		-32	133	447	0.034557	0.863248
-100	201	379	0.11655	0.668874		-31	132	448	0.033441	0.866953
-99	200	380	0.11525	0.671096		-30	131	449	0.032328	0.87069
-98	199	381	0.113953	0.673333		-29	130	450	0.031216	0.874459
-97	198	382	0.11266	0.675585		-28	129	451	0.030108	0.878261
-96	197	383	0.111369	0.677852		-27	128	452	0.029001	0.882096
-95	196	384	0.110081	0.680135		-26	127	453	0.027897	0.885965
-94	195	385	0.108796	0.682432		-25	126	454	0.026795	0.889868

-93	194	386	0.107514	0.684746		-24	125	455	0.025696	0.893805
-92	193	387	0.106236	0.687075		-23	124	456	0.024599	0.897778
-91	192	388	0.10496	0.68942		-22	123	457	0.023504	0.901786
-90	191	389	0.103687	0.691781		-21	122	458	0.022412	0.90583
-89	190	390	0.102417	0.694158		-20	121	459	0.021322	0.90991
-88	189	391	0.101149	0.696552		-19	120	460	0.020234	0.914027
-87	188	392	0.099885	0.698962		-18	119	461	0.019149	0.918182
-86	187	393	0.098624	0.701389		-17	118	462	0.018066	0.922374
-85	186	394	0.097365	0.703833		-16	117	463	0.016985	0.926606
-84	185	395	0.09611	0.706294		-15	116	464	0.015907	0.930876
-83	184	396	0.094857	0.708772		-14	115	465	0.014831	0.935185
-82	183	397	0.093607	0.711268		-13	114	466	0.013757	0.939535
-81	182	398	0.09236	0.713781		-12	113	467	0.012685	0.943925
-80	181	399	0.091116	0.716312		-11	112	468	0.011616	0.948357
-79	180	400	0.089875	0.718861		-10	111	469	0.010549	0.95283
-78	179	401	0.088636	0.721429		-9	110	470	0.009484	0.957346
-77	178	402	0.087401	0.724014		-8	109	471	0.008421	0.961905
-76	177	403	0.086168	0.726619		-7	108	472	0.007361	0.966507
-75	176	404	0.084938	0.729242		-6	107	473	0.006303	0.971154
-74	175	405	0.08371	0.731884		-5	106	474	0.005247	0.975845
-73	174	406	0.082486	0.734545		-4	105	475	0.004193	0.980583
-72	173	407	0.081264	0.737226		-3	104	476	0.003141	0.985366
-71	172	408	0.080045	0.739927		-2	103	477	0.002092	0.990196
						-1	102	478	0.001045	0.995074

Similarly, Table 12 and 13 show all the feasible pairs of sensitivity and specificity for cases and controls, respectively, for Example 2, assuming only over-misclassification exist. Over-

misclassification means the observed number of "Use" is higher than the true number, as indicated by a positive $m_{(1)j}$, j=1,0). Again, values of (1-Sp) instead of Sp are listed.

Table 12 Feasible 1-Sp and Se for over-misclassification of Example 2, Case.

$m_{(1)1}$	Use	No use	1-Sp	Se		$m_{(1)1}$	Use	No use	1-Sp	Se
+1	121	443	0.00113	0.995885		+54	68	496	0.057569	0.715789
+2	120	444	0.002257	0.991736		+55	67	497	0.058573	0.708995
+3	119	445	0.003382	0.987552		+56	66	498	0.059574	0.702128
+4	118	446	0.004505	0.983333		+57	65	499	0.060574	0.695187
+5	117	447	0.005624	0.979079		+58	64	500	0.061571	0.688172
+6	116	448	0.006742	0.97479		+59	63	501	0.062566	0.681081
+7	115	449	0.007856	0.970464		+60	62	502	0.063559	0.673913
+8	114	450	0.008969	0.966102		+61	61	503	0.06455	0.666667
+9	113	451	0.010078	0.961702		+62	60	504	0.065539	0.659341
+10	112	452	0.011186	0.957265		+63	59	505	0.066526	0.651934
+11	111	453	0.012291	0.95279		+64	58	506	0.067511	0.644444
+12	110	454	0.013393	0.948276		+65	57	507	0.068493	0.636872
+13	109	455	0.014493	0.943723		+66	56	508	0.069474	0.629213
+14	108	456	0.01559	0.93913		+67	55	509	0.070452	0.621469
+15	107	457	0.016685	0.934498		+68	54	510	0.071429	0.613636
+16	106	458	0.017778	0.929825		+69	53	511	0.072403	0.605714
+17	105	459	0.018868	0.92511		+70	52	512	0.073375	0.597701
+18	104	460	0.019956	0.920354		+71	51	513	0.074346	0.589595
+19	103	461	0.021041	0.915556		+72	50	514	0.075314	0.581395
+20	102	462	0.022124	0.910714		+73	49	515	0.07628	0.573099
+21	101	463	0.023204	0.90583		+74	48	516	0.077244	0.564706
+22	100	464	0.024283	0.900901		+75	47	517	0.078206	0.556213
+23	99	465	0.025358	0.895928		+76	46	518	0.079167	0.547619
+24	98	466	0.026432	0.890909		+77	45	519	0.080125	0.538922

+25	97	467	0.027503	0.885845		+78	44	520	0.081081	0.53012
+26	96	468	0.028571	0.880734		+79	43	521	0.082035	0.521212
+27	95	469	0.029638	0.875576		+80	42	522	0.082988	0.512195
+28	94	470	0.030702	0.87037		+81	41	523	0.083938	0.503067
+29	93	471	0.031763	0.865116		+82	40	524	0.084886	0.493827
+30	92	472	0.032823	0.859813		+83	39	525	0.085832	0.484472
+31	91	473	0.03388	0.85446		+84	38	526	0.086777	0.475
+32	90	474	0.034934	0.849057		+85	37	527	0.087719	0.465409
+33	89	475	0.035987	0.843602		+86	36	528	0.08866	0.455696
+34	88	476	0.037037	0.838095		+87	35	529	0.089598	0.44586
+35	87	477	0.038085	0.832536		+88	34	530	0.090535	0.435897
+36	86	478	0.03913	0.826923		+89	33	531	0.09147	0.425806
+37	85	479	0.040174	0.821256		+90	32	532	0.092402	0.415584
+38	84	480	0.041215	0.815534		+91	31	533	0.093333	0.405229
+39	83	481	0.042254	0.809756		+92	30	534	0.094262	0.394737
+40	82	482	0.04329	0.803922		+93	29	535	0.095189	0.384106
+41	81	483	0.044324	0.79803		+94	28	536	0.096115	0.373333
+42	80	484	0.045356	0.792079		+95	27	537	0.097038	0.362416
+43	79	485	0.046386	0.78607		+96	26	538	0.097959	0.351351
+44	78	486	0.047414	0.78		+97	25	539	0.098879	0.340136
+45	77	487	0.048439	0.773869		+98	24	540	0.099796	0.328767
+46	76	488	0.049462	0.767677		+99	23	541	0.100712	0.317241
+47	75	489	0.050483	0.761421		+100	22	542	0.101626	0.305556
+48	74	490	0.051502	0.755102		+101	21	543	0.102538	0.293706
+49	73	491	0.052519	0.748718		+102	20	544	0.103448	0.28169
+50	72	492	0.053533	0.742268		+103	19	545	0.104357	0.269504
+51	71	493	0.054545	0.735751		+104	18	546	0.105263	0.257143
+52	70	494	0.055556	0.729167		+105	17	547	0.106168	0.244604
+53	69	495	0.056564	0.722513		+106	16	548	0.107071	0.231884

Table 13 Feasible 1-Sp and Se for over-misclassification of Example 2, Control.

$m_{(1)0}$	Use	No use	1-Sp	Se		$m_{(1)0}$	Use	No use	1-Sp	Se
+1	100	480	0.001043	0.995025		+46	55	525	0.045817	0.705128
+2	99	481	0.002083	0.99		+47	54	526	0.046766	0.696774
+3	98	482	0.003122	0.984925		+48	53	527	0.047714	0.688312
+4	97	483	0.004158	0.979798		+49	52	528	0.048659	0.679739
+5	96	484	0.005192	0.974619		+50	51	529	0.049603	0.671053
+6	95	485	0.006224	0.969388		+51	50	530	0.050545	0.662252
+7	94	486	0.007254	0.964103		+52	49	531	0.051485	0.653333
+8	93	487	0.008282	0.958763		+53	48	532	0.052423	0.644295
+9	92	488	0.009307	0.953368		+54	47	533	0.05336	0.635135
+10	91	489	0.010331	0.947917		+55	46	534	0.054294	0.62585
+11	90	490	0.011352	0.942408		+56	45	535	0.055227	0.616438
+12	89	491	0.012371	0.936842		+57	44	536	0.056158	0.606897
+13	88	492	0.013388	0.931217		+58	43	537	0.057087	0.597222
+14	87	493	0.014403	0.925532		+59	42	538	0.058014	0.587413
+15	86	494	0.015416	0.919786		+60	41	539	0.058939	0.577465
+16	85	495	0.016427	0.913978		+61	40	540	0.059863	0.567376
+17	84	496	0.017436	0.908108		+62	39	541	0.060784	0.557143
+18	83	497	0.018443	0.902174		+63	38	542	0.061704	0.546763
+19	82	498	0.019447	0.896175		+64	37	543	0.062622	0.536232
+20	81	499	0.02045	0.89011		+65	36	544	0.063539	0.525547
+21	80	500	0.02145	0.883978		+66	35	545	0.064453	0.514706
+22	79	501	0.022449	0.877778		+67	34	546	0.065366	0.503704
+23	78	502	0.023445	0.871508		+68	33	547	0.066277	0.492537
+24	77	503	0.02444	0.865169		+69	32	548	0.067186	0.481203
+25	76	504	0.025432	0.858757		+70	31	549	0.068093	0.469697
+26	75	505	0.026423	0.852273		+71	30	550	0.068999	0.458015
+27	74	506	0.027411	0.845714		+72	29	551	0.069903	0.446154
+28	73	507	0.028398	0.83908		+73	28	552	0.070805	0.434109

+29	72	508	0.029382	0.83237		+74	27	553	0.071705	0.421875
+30	71	509	0.030364	0.825581		+75	26	554	0.072604	0.409449
+31	70	510	0.031345	0.818713		+76	25	555	0.073501	0.396825
+32	69	511	0.032323	0.811765		+77	24	556	0.074396	0.384
+33	68	512	0.0333	0.804734		+78	23	557	0.07529	0.370968
+34	67	513	0.034274	0.797619		+79	22	558	0.076181	0.357724
+35	66	514	0.035247	0.790419		+80	21	559	0.077071	0.344262
+36	65	515	0.036217	0.783133		+81	20	560	0.07796	0.330579
+37	64	516	0.037186	0.775758		+82	19	561	0.078846	0.316667
+38	63	517	0.038153	0.768293		+83	18	562	0.079731	0.302521
+39	62	518	0.039117	0.760736		+84	17	563	0.080614	0.288136
+40	61	519	0.04008	0.753086		+85	16	564	0.081496	0.273504
+41	60	520	0.041041	0.745342		+86	15	565	0.082375	0.258621
+42	59	521	0.042	0.7375		+87	14	566	0.083254	0.243478
+43	58	522	0.042957	0.72956		+88	13	567	0.08413	0.22807
+44	57	523	0.043912	0.721519		+89	12	568	0.085005	0.212389
+45	56	524	0.044865	0.713376		+90	11	569	0.085878	0.196429
						+91	10	570	0.086749	0.18018

Table 14 and 15 list all the calculated feasible pairs of sensitivity (Se) and specificity (Sp) for cases and controls, respectively, for Example 3, where only under-misclassification for exposed subjects ("User") is assumed to exist. Under-misclassification means the observed number of exposed subjects is less than the true number (as indicated by a negative $m_{(1)j}$, j=1,0). Values of (1-Sp) instead of Sp are listed.

Table 14 Feasible 1-Sp and Se for under-misclassification of Example 3, Case.

$m_{(1)1}$	User	Nonuser	1-Sp	Se		$m_{(1)1}$	User	Nonuser	1-Sp	Se
-59	114	69	0.299492	0.650888		-29	84	99	0.127753	0.791367
-58	113	70	0.292929	0.654762		-28	83	100	0.122807	0.797101
-57	112	71	0.286432	0.658683		-27	82	101	0.117904	0.80292
-56	111	72	0.28	0.662651		-26	81	102	0.113043	0.808824
-55	110	73	0.273632	0.666667		-25	80	103	0.108225	0.814815
-54	109	74	0.267327	0.670732		-24	79	104	0.103448	0.820896
-53	108	75	0.261084	0.674847		-23	78	105	0.098712	0.827068
-52	107	76	0.254902	0.679012		-22	77	106	0.094017	0.833333
-51	106	77	0.24878	0.68323		-21	76	107	0.089362	0.839695
-50	105	78	0.242718	0.6875		-20	75	108	0.084746	0.846154
-49	104	79	0.236715	0.691824		-19	74	109	0.080169	0.852713
-48	103	80	0.230769	0.696203		-18	73	110	0.07563	0.859375
-47	102	81	0.22488	0.700637		-17	72	111	0.07113	0.866142
-46	101	82	0.219048	0.705128		-16	71	112	0.066667	0.873016
-45	100	83	0.21327	0.709677		-15	70	113	0.062241	0.88
-44	99	84	0.207547	0.714286		-14	69	114	0.057851	0.887097
-43	98	85	0.201878	0.718954		-13	68	115	0.053498	0.894309
-42	97	86	0.196262	0.723684		-12	67	116	0.04918	0.901639
-41	96	87	0.190698	0.728477		-11	66	117	0.044898	0.909091
-40	95	88	0.185185	0.733333		-10	65	118	0.04065	0.916667
-39	94	89	0.179724	0.738255		-9	64	119	0.036437	0.92437
-38	93	90	0.174312	0.743243		-8	63	120	0.032258	0.932203
-37	92	91	0.16895	0.748299		-7	62	121	0.028112	0.940171
-36	91	92	0.163636	0.753425		-6	61	122	0.024	0.948276
-35	90	93	0.158371	0.758621		-5	60	123	0.01992	0.956522
-34	89	94	0.153153	0.763889		-4	59	124	0.015873	0.964912
-33	88	95	0.147982	0.769231		-3	58	125	0.011858	0.973451
-32	87	96	0.142857	0.774648		-2	57	126	0.007874	0.982143
-31	86	97	0.137778	0.780142		-1	56	127	0.003922	0.990991
-30	85	98	0.132743	0.785714						

Table 15 Feasible 1-Sp and Se for under-misclassification of Example 3, Control

$m_{(1)0}$	User	Nonuser	1-Sp	Se		$m_{(1)0}$	User	Nonuser	1-Sp	Se
-29	48	135	0.09699	0.567164		-14	33	150	0.044586	0.730769
-28	47	136	0.093333	0.575758		-13	32	151	0.04127	0.745098
-27	46	137	0.089701	0.584615		-12	31	152	0.037975	0.76
-26	45	138	0.086093	0.59375		-11	30	153	0.0347	0.77551
-25	44	139	0.082508	0.603175		-10	29	154	0.031447	0.791667
-24	43	140	0.078947	0.612903		-9	28	155	0.028213	0.808511
-23	42	141	0.07541	0.622951		-8	27	156	0.025	0.826087
-22	41	142	0.071895	0.633333		-7	26	157	0.021807	0.844444
-21	40	143	0.068404	0.644068		-6	25	158	0.018634	0.863636
-20	39	144	0.064935	0.655172		-5	24	159	0.01548	0.883721
-19	38	145	0.061489	0.666667		-4	23	160	0.012346	0.904762
-18	37	146	0.058065	0.678571		-3	22	161	0.009231	0.926829
-17	36	147	0.054662	0.690909		-2	21	162	0.006135	0.95
-16	35	148	0.051282	0.703704		-1	20	163	0.003058	0.974359
-15	34	149	0.047923	0.716981						

Table 16 and 17 list all the calculated feasible pairs of sensitivity (Se) and specificity (Sp) for cases and controls, respectively, for Example 3, where only over-misclassification for exposed subjects ("User") is assumed to exist. Over-misclassification means the observed number of exposed subjects is more than the true number (as indicated by a positive $m_{(1)j}$, j=1,0). For convenience, again values of (1-Sp) instead of Sp are listed.

Table 16 Feasible 1-Sp and Se for over-misclassification of Example 3, Case.

$m_{(1)1}$	User	Nonuser	1-Sp	Se		$m_{(1)1}$	User	Nonuser	1-Sp	Se
+1	54	129	0.003891	0.990826		+24	31	152	0.085714	0.72093
+2	53	130	0.007752	0.981481		+25	30	153	0.088968	0.705882

+3	52	131	0.011583	0.971963		+26	29	154	0.092199	0.690476
+4	51	132	0.015385	0.962264		+27	28	155	0.095406	0.674699
+5	50	133	0.019157	0.952381		+28	27	156	0.098592	0.658537
+6	49	134	0.022901	0.942308		+29	26	157	0.101754	0.641975
+7	48	135	0.026616	0.932039		+30	25	158	0.104895	0.625
+8	47	136	0.030303	0.921569		+31	24	159	0.108014	0.607595
+9	46	137	0.033962	0.910891		+32	23	160	0.111111	0.589744
+10	45	138	0.037594	0.9		+33	22	161	0.114187	0.571429
+11	44	139	0.041199	0.888889		+34	21	162	0.117241	0.552632
+12	43	140	0.044776	0.877551		+35	20	163	0.120275	0.533333
+13	42	141	0.048327	0.865979		+36	19	164	0.123288	0.513514
+14	41	142	0.051852	0.854167		+37	18	165	0.12628	0.493151
+15	40	143	0.055351	0.842105		+38	17	166	0.129252	0.472222
+16	39	144	0.058824	0.829787		+39	16	167	0.132203	0.450704
+17	38	145	0.062271	0.817204		+40	15	168	0.135135	0.428571
+18	37	146	0.065693	0.804348		+41	14	169	0.138047	0.405797
+19	36	147	0.069091	0.791209		+42	13	170	0.14094	0.382353
+20	35	148	0.072464	0.777778		+43	12	171	0.143813	0.358209
+21	34	149	0.075812	0.764045		+44	11	172	0.146667	0.333333
+22	33	150	0.079137	0.75		+45	10	173	0.149502	0.307692
+23	32	151	0.082437	0.735632						

Table 17 Feasible 1-Sp and Se for over-misclassification of Example 3, Control.

$m_{(1)0}$	User	Nonuser	1-Sp	Se		$m_{(1)0}$	User	Nonuser	1-Sp	Se
+1	18	165	0.00304	0.972973		+10	9	174	0.029586	0.642857
+2	17	166	0.006061	0.944444		+11	8	175	0.032448	0.592593
+3	16	167	0.009063	0.914286		+12	7	176	0.035294	0.538462
+4	15	168	0.012048	0.882353		+13	6	177	0.038123	0.48
+5	14	169	0.015015	0.848485		+14	5	178	0.040936	0.416667
+6	13	170	0.017964	0.8125		+15	4	179	0.043732	0.347826

+7	12	171	0.020896	0.774194		+16	3	180	0.046512	0.272727
+8	11	172	0.02381	0.733333		+17	2	181	0.049275	0.190476
+9	10	173	0.026706	0.689655						

3.2 The empirical EOC Curves

Figures 2 through 7 shows, respectively, the empirical EOC curves for under-misclassification and over-misclassification, for the three examples, using the feasible sensitivity and specificity values calculated in Table 6 through 17. Each figure includes both case and control for visual comparison.

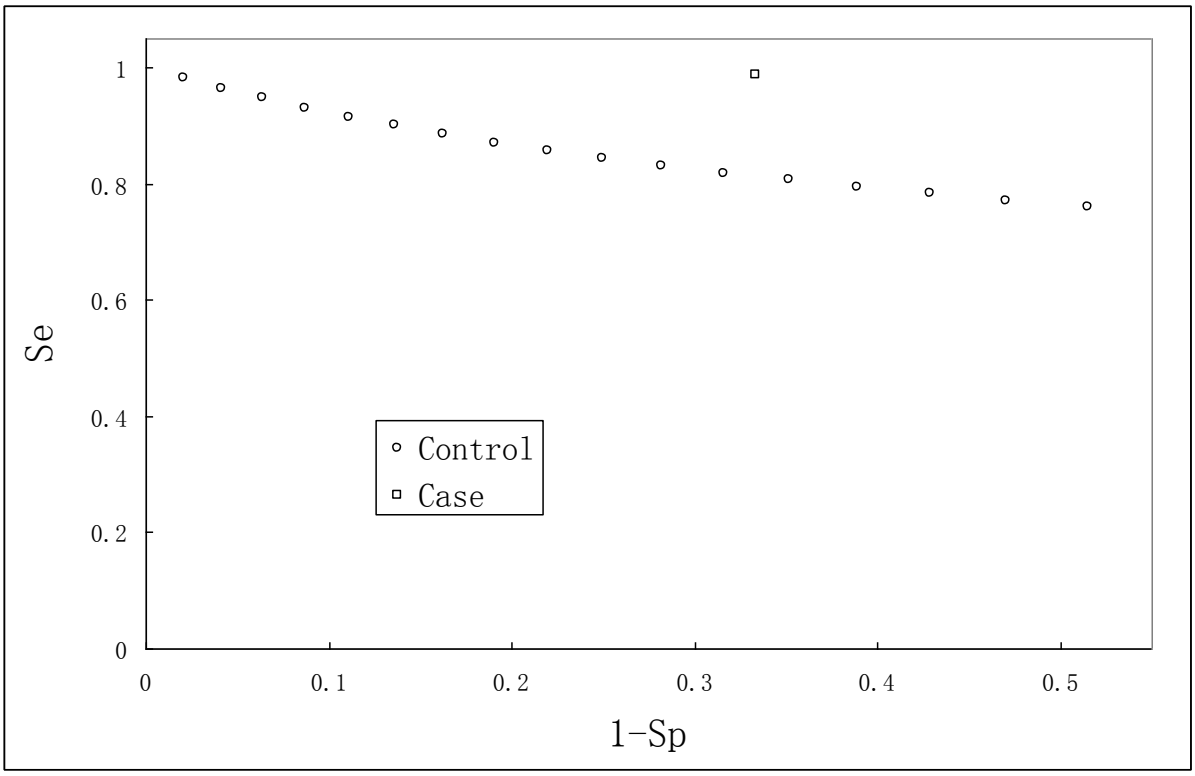

Fig. 1 EOC curves of Example 1, under-misclassification

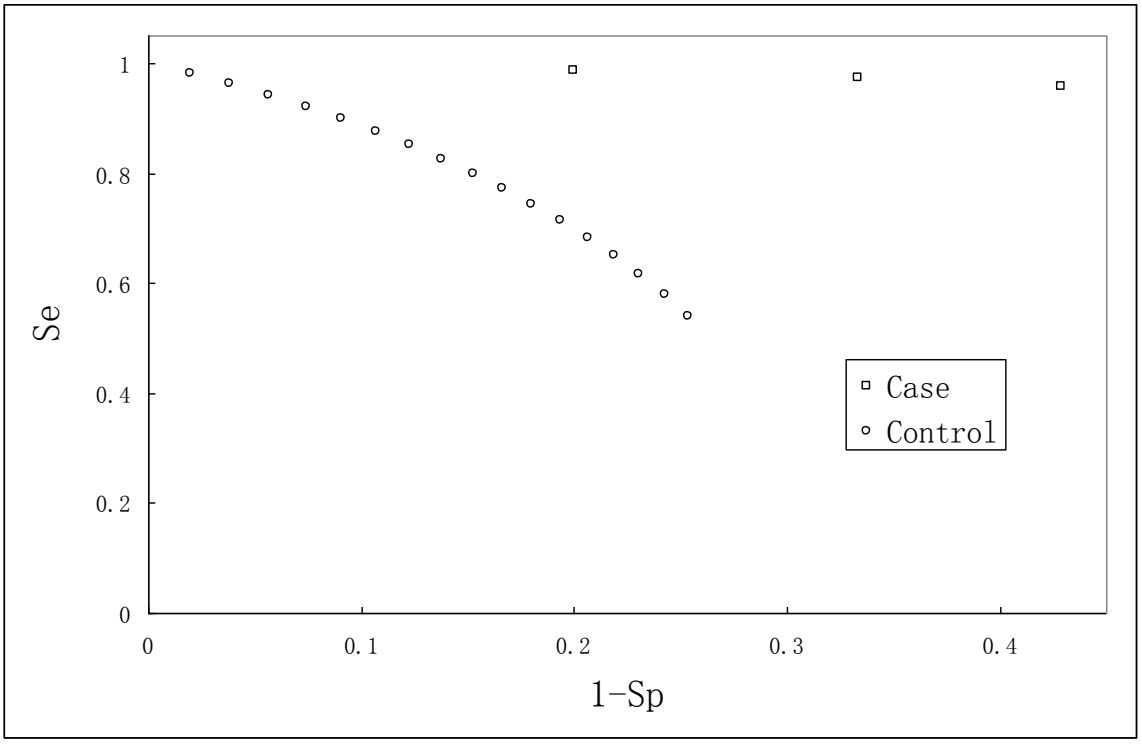

Fig. 2 EOC curves of Example 1, over-misclassification

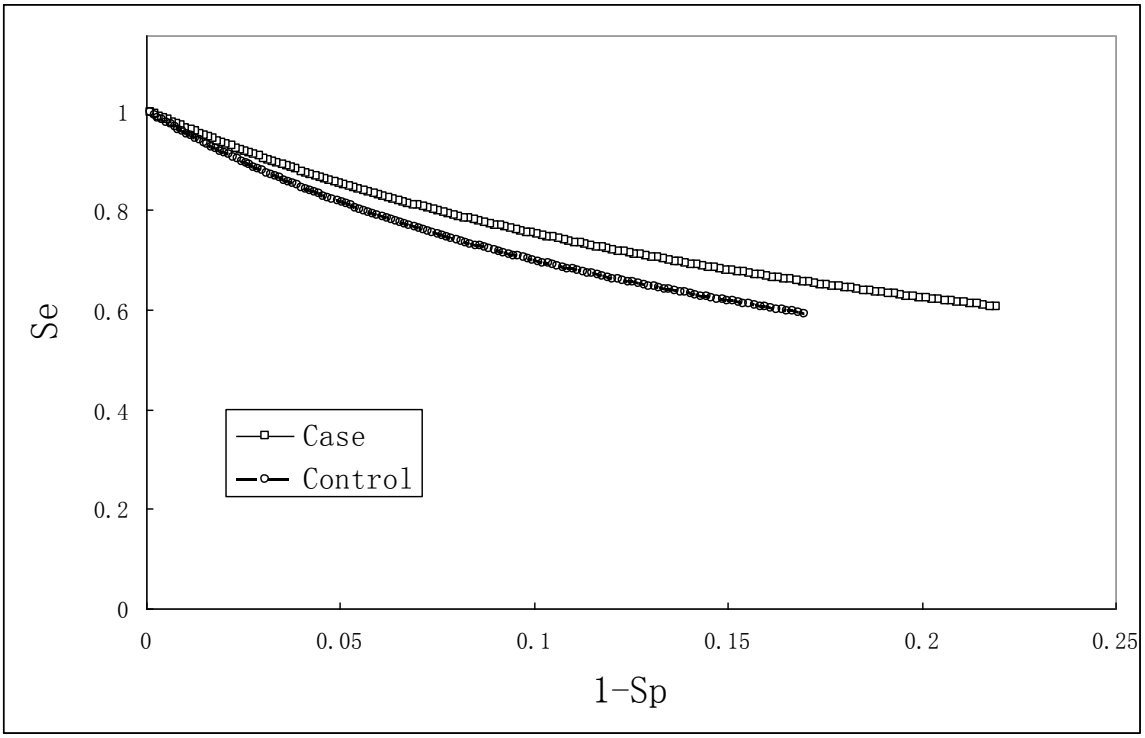

Fig. 3 EOC curves of Example 2, under-misclassification

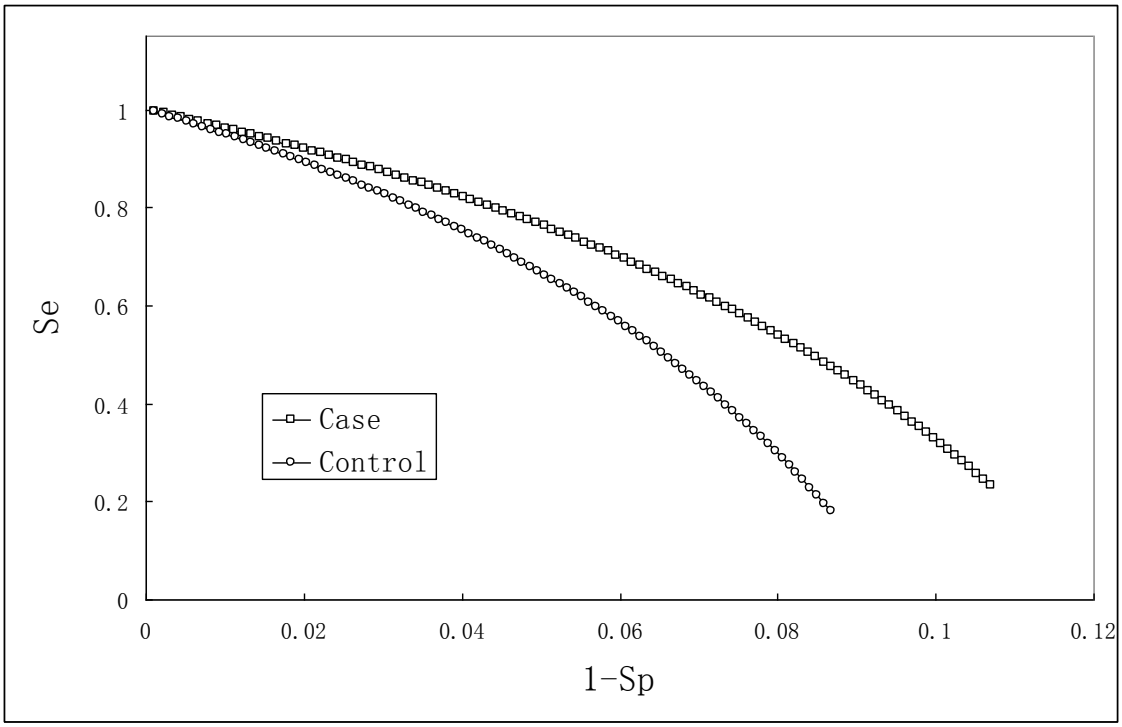

Fig. 4 EOC curves of Example 2, over-misclassification

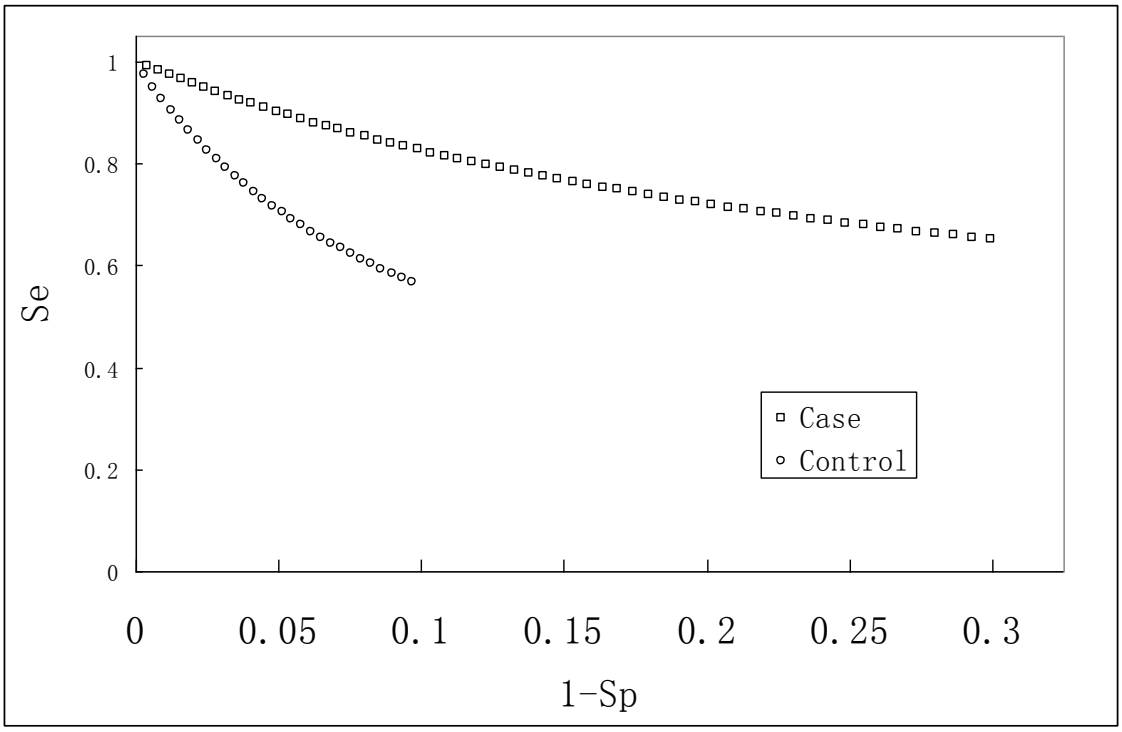

Fig. 5 EOC curves of Example 3, under-misclassification

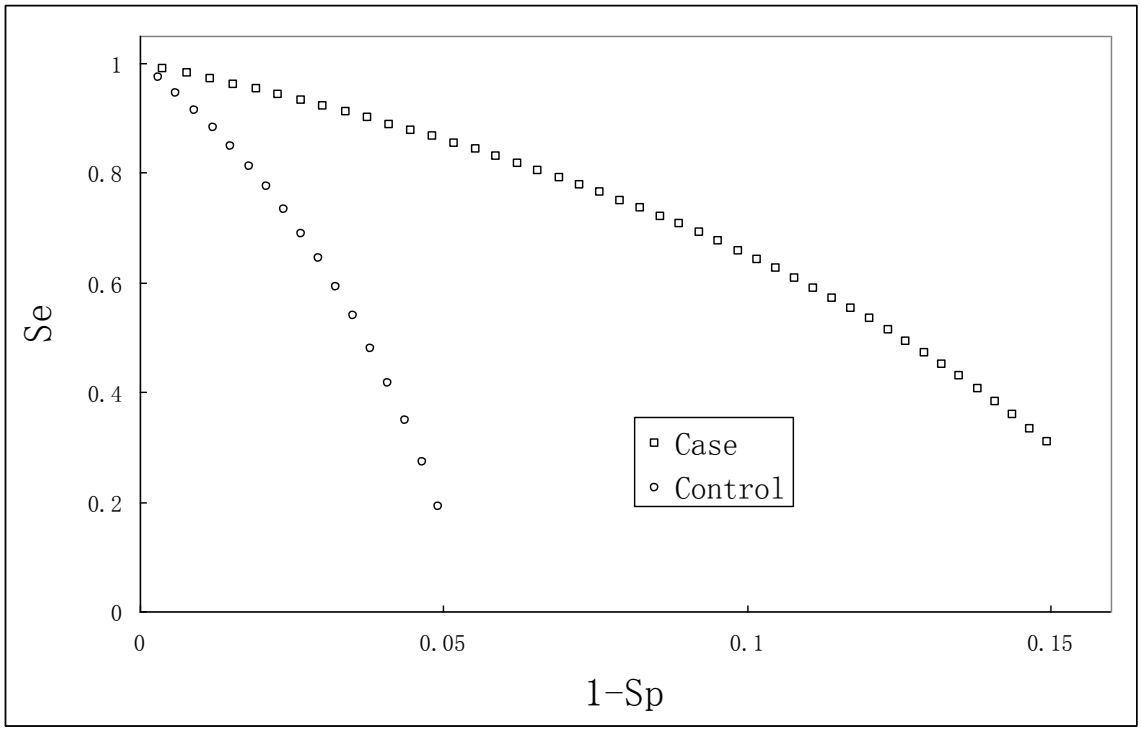

Fig. 6 EOC curves of Example 3, over-misclassification

3.3 Hypothesis Testing

3.3.1 Using Method A

The statistic calculated using formula 11, and the corresponding standard errors for both under- and over-misclassified situations are summarized in Tables 18 through 20, for all three examples, along with the standardized z_{MW}-statistic for testing the hypothesis (formula 12).

Because the case data of example 1 under-misclassification has only one point, this calculation could not be applied to it. The rest five sets results are very scattered. Two of them show very small *p*-values, suggesting we reject the null hypothesis, while other three show the opposite.

Table 18 The Mann-Whiney Statistic for Example 1

	Misclassification Type	
	Under	Over
θ_{MW}	_	0.9412
$SE(\theta_{MW})$	_	0.1852
z_{MW}	_	2.382
p-value	_	0.00861

Table 19 The Mann-Whiney Statistic for Example 2

	Misclassification Type	
	Under	Over
θ_{MW}	0.5271	0.5189
$SE(\theta_{MW})$	0.0336	0.0414
z_{MW}	0.8065	0.4565
p-value	0.2100	0.3240

Table 20 The Mann-Whiney Statistic for Example 3

	Misclassification Type	
	Under	Over
θ_{MW}	0.6596	0.5608
$SE(\theta_{MW})$	0.0658	0.0828
z_{MW}	2.426	0.7343
p-value	0.00763	0.2314

3.3.2 Using Method B (the Graphical Method)

For under and over misclassification of each example, the original data from case and control are combined to make a new data set. A quadratic curve is then fitted using least square fitting and the mixed data set. Then numbers of points above the fitted curve from case, v, and the total number of points above the fitted curve from both case and control, m, are counted and used to calculate *p*-values using formula 13.

Figures 8 through 13 show the fitted curves along with the empirical EOC curves for examples 1 to 3, under- and over-misclassification, respectively. Tables 21 to 23 summarize the calculated parameters and *p*-values for the three examples. From the figures we can see that, except the one for example 1 over-misclassfication, in all other five figures the empirical EOC curves of case and control show distinct behavior. All or most of the points from case are above the fitted curve, while most points from control are below the fitted curve. This characteristic is

readily confirmed by the calculated *p*-values, which are extremely small for examples 2 and 3, and is fairly small, too, for example 1 under-misclassification.

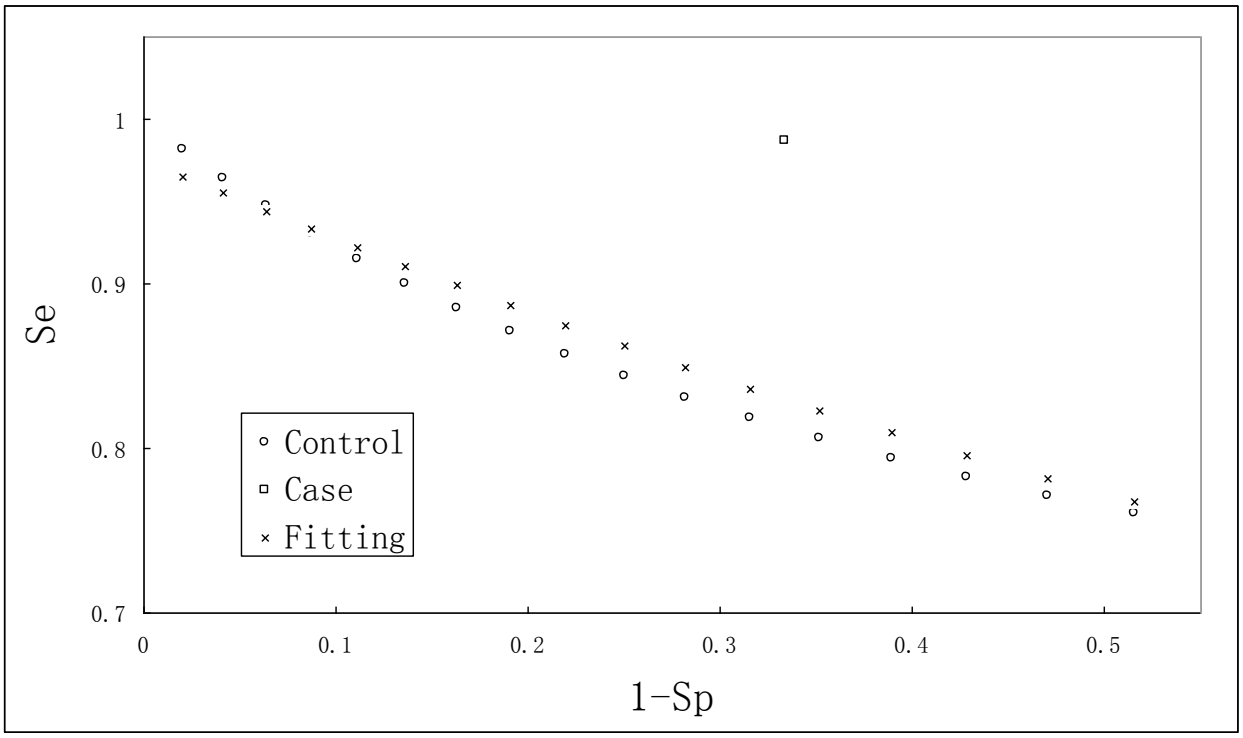

Fig. 7 Fitting results for Example 1, under-misclassification

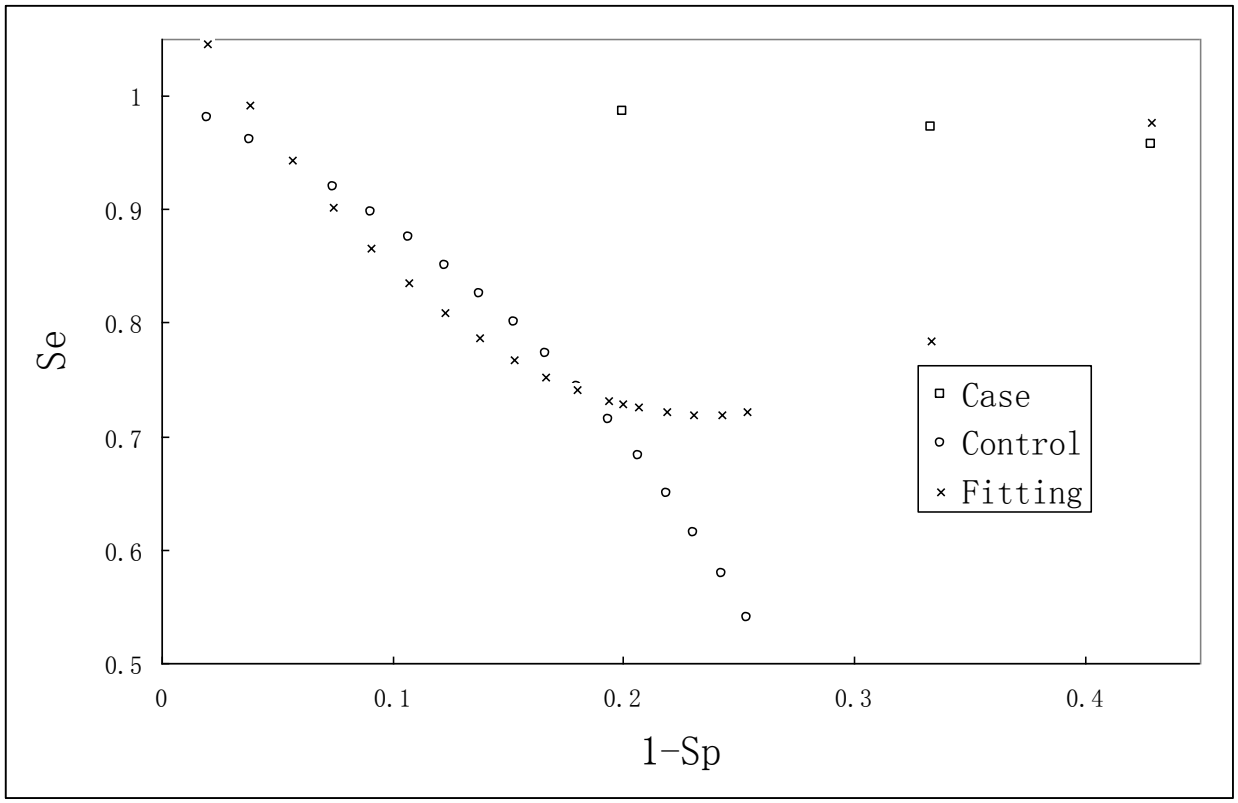

Fig. 8 Fitting results for Example 1, over-misclassification

Table 21 The Graphical Method results for Example 1

	Misclassification Type	
	Under	Over
m_{K_0}	1	3
m_{K_1}	17	17
v_0	1	2
v_1	3	6
p-value	0.22	0.8485

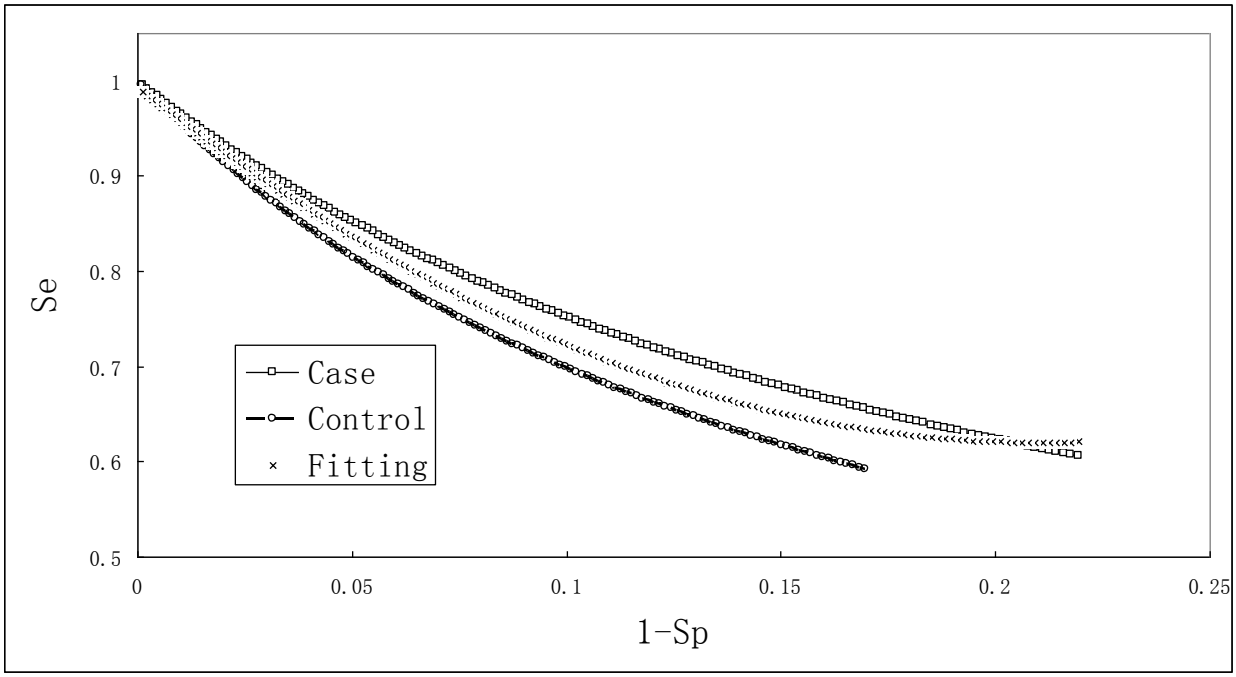

Fig. 9 Fitting results for Example 2, under-misclassification

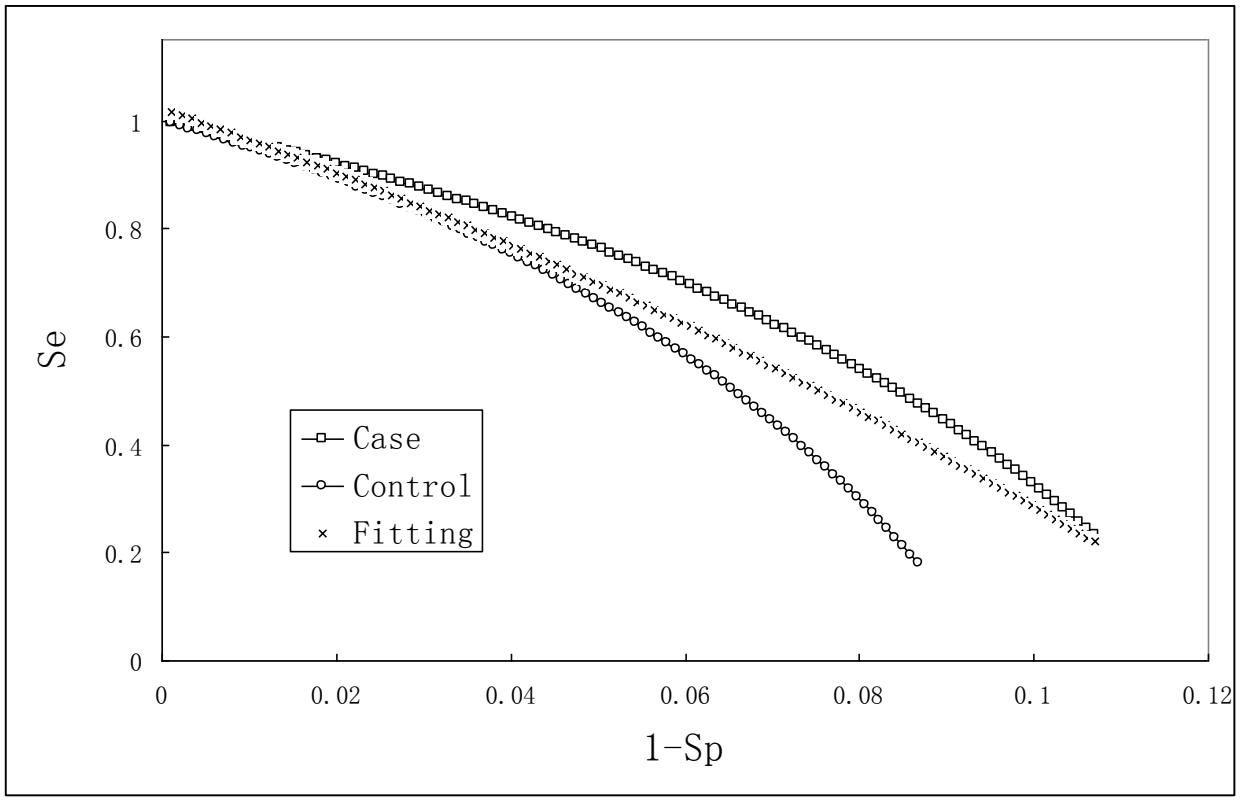

Fig. 10 Fitting results for Example 2, over-misclassification

Table 22 The Graphical Method results for Example 2

	Misclassification Type	
	Under	Over
m_{K_0}	159	106
m_{K_1}	139	91
v_0	150	100
v_1	3	3
p-value	0.0000	0.0000

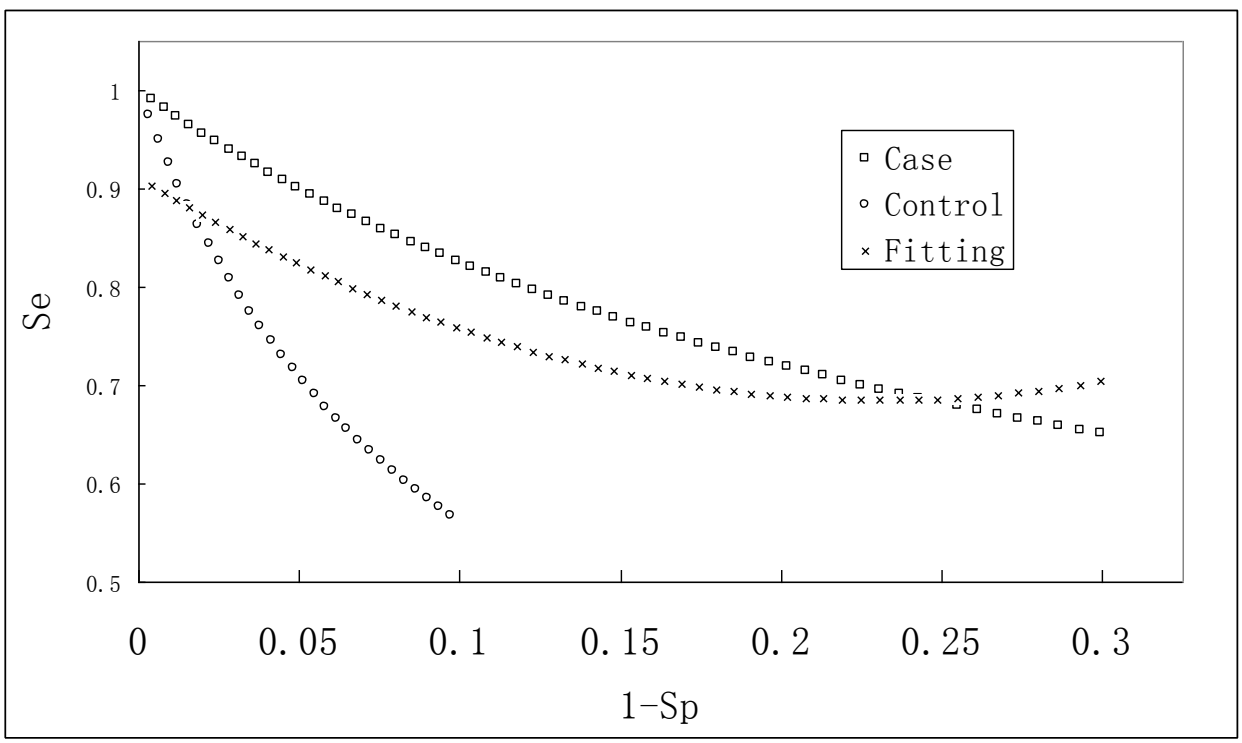

Fig. 11 Fitting results for Example 3, under-misclassification

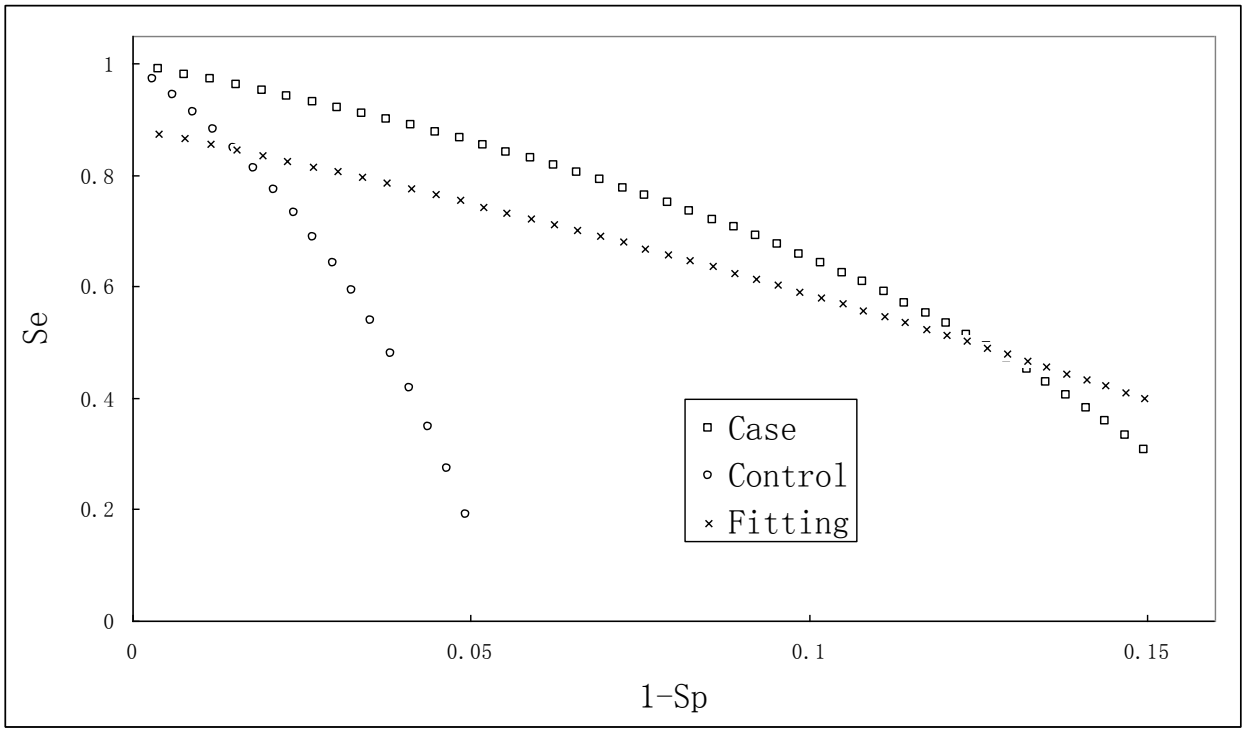

Fig. 12 Fitting results for Example 3, over-misclassification

Table 23 The Graphical Method results for Example 3

	Misclassification Type	
	Under	Over
m_{K_0}	59	45
m_{K_1}	29	17
v_0	50	37
v_1	5	5
p-value	0.0000	0.0000

It should be noted that the *p*-values calculated using formula 13 is one-sided. However, as the *p*-values are extremely small (except example 1), the two-sided test *p*-values are still close to

zero. Therefore for example 2 and 3, we must reject the null hypothesis, and conclude that $EOC_1 \neq EOC_0$. Because the *EOC* curves of example 1 (case) have too few points, the results for that example is not conclusive.

CHAPTER 4 CONCLUTION AND DISCUSSION

As a summary, a new method, called EOC curve, has been developed for the first time to test the non-differential assumption in case – control studies when misclassifications exist. Three examples from literature have been used to demonstrate this new method. Two procedures have been employed to test the null hypothesis, i.e., whether the non-differential assumption is valid.

First of all, by looking at the empirical EOC curves (figures 2 to 7) we can make some preliminary conclusion. Because none of the EOC curves for cases and controls in all examples cross each other, it seems intuitively clear that the assumption of non-differential misclassification is unlikely valid.

However, the results of hypothesis testing using the two procedures are not consistent with each other. One thing is clear that all the data sets have to have more than a few points, or the procedures will not work properly. This is why the result for example 1 is not conclusive.

Method A only rejects the null hypothesis for the under-misclassification in example 3. As a side remark on Method A, it is a nonparametric method which suffers the lack of power. Furthermore, it is possible two EOC curves are different but have the same area under the curve if two curves cross each other.

To the contrary of Method A, Method B rejects the null hypothesis for examples 2 and 3, clearly showing that the non-differential assumption of misclassification is not valid in these

case-control studies. As a result, estimation based on this assumption may be completely wrong. It is worth pointing out that this result is consistent with the findings in example 2 using validation data (Greenland, 1988; also see Chapter 1). As a conclusion, the proposed method, i.e. using the EOC curve, is able to test if the misclassification is differential or non-differential, and Method B is recommended to be employed for testing the assumption of non-differential misclassification in case-control studies.

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., and Garcia, R. (1979). Endometrial Cancer and Estrogen Use, Report of a Large Case-Control Study. New England Journal of Medicine, 300, 9-13.

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

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

Chen, T. T. (1989). A review of methods for misclassified categorical data in epidemiology. Stat Med, 81, 1095-1106.

Chu, H., Wang, Z., Cole, S. R., and Greenland, S. (2006). Sensitivity analysis of misclassification: A graphical and a Bayesian approach. Ann Epidemiol, 16, 834-841.

Drews, C., Kraus, J. and Greenland, S (1987). The magnitude of recall error in a case-control study of sudden infant death syndrome. American Journal of Epidemiology, 126, 745 (abstract).

Espeland, M. A., Hui, S. L. (1987). A General Approach to Analyzing Epidemiologic Data That Contain Misclassification Errors. Biometrics 43, 1001-1012.

Greenland, S. (1988). Variance estimation for epidemiologic effect estimates under misclassification. Stat Med, 7, 745-757.

Lee, T-S. (2009). Bias-adjusted exposure odds ratio for misclassified data. The Internet Journal of Epidemiology, 6, 1-19.

Lusted, L. B. (1971). Signal detectability and medical decision-making, Science, 171, 1217-1219.

Nelson, L. M., Longstreth, W. T., Koepsell, T. D., and van Belle, G. (1990). Proxy respondents in epidemiologic research. Epidemiol Rev, 12, 71-86.

Rocca, W. A., Fratiglioni, L., Bracco, L., Pedone, D., Groppi, C., and Schoenberg, B. S. (1986). The use of surrogate respondents to obtain questionnaire data in case-control studies of neurological diseases. J Chron Dis, 39, 907-912.

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

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

Sprent, P. and Smeeton, N. C. (2007). Applied Nonparametric Statistical Methods 4th Edition. Boca Raton, FL: Chapman & Hall/CRC.

Tsutakama, R. K. and Hewett, J. E. (1977). Quick test for comparing two populations with bivariate data. Biometrics, 33, 215-219.

Walker, S. D. and 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., Obuchowski, N. A. and McClish, D. K. (2002). Statistical Methods in Diagnostic Medicine. New York: Wiley.

APPENDIX

R-code for Method A (the Mann-Whiney statistic):

```
# Mann-Whiney Statistic w - n0 not eg n1

#Read in data

x0 <- read.table("C:\\Users\\DAD\\GSU\\Thesis\\Dr. Lee\\Calculation B\\Ex1UpperCase.txt")

x1 <- read.table("C:\\Users\\DAD\\GSU\\Thesis\\Dr. Lee\\Calculation B\\Ex1UpperControl.txt")



#Number of sample pairs

n0<-length(x0[,1])

n1<-length(x1[,1])


s<-0

sum<-0

w=0


#Compare all possible pairs

for(i in 1:n0){

  for(j in 1:n1){

    if (x0[i,1]>x1[j,1]) {s<-1}

        else if (x0[i,1]<x1[j,1]) {s<-0}

            else {s<-0.5}

sum<-sum+s
```

```
  }
}

#The Mann-Whiney Statistic
w<-sum/n0/n1

print(w)

#Optional: Varance of w
Q1<-w/(2-w)
Q2<-2*w*w/(1+w)
var<-(w*(1-w)+(n1-1)*(Q1-w*w)+(n0-1)*(Q2-w*w))/n0/n1

print(var)
```

R-code for Method B (the Graphic Method):

```
#Read Mixed Data
data <- read.table("C:\\Users\\DAD\\GSU\\Thesis\\Dr. Lee\\Calculation C\\Ex1UnderMix.txt")

#Sort and plot mixed data
(ii<-order(x<-data[,1], y<-data[,2]))
data.sort<-rbind(x,y)[,ii]
```

```
plot(x,y, xlim=c(0,1), ylim=c(0,1))

#fitting mixed data
fit2 <- lm( y~poly(x,2) )
xx <- sort(x)
lines(xx, predict(fit2, data.frame(x=xx)), col='red')

#Read case and control data
data0 <- read.table("C:\\Users\\DAD\\GSU\\Thesis\\Dr. Lee\\Calculation C\\Ex1UnderCase.txt")
data1 <- read.table("C:\\Users\\DAD\\GSU\\Thesis\\Dr. Lee\\Calculation
C\\Ex1UnderControl.txt")

x0<-data0[,1]
y0<-data0[,2]
y20<-predict(fit2, data.frame(x=x0))
n0<-length(x0)

x1<-data1[,1]
y1<-data1[,2]
y21<-predict(fit2, data.frame(x=x1))
n1<-length(x1)

n<-n0+n1
```

```
v0<-0
v1<-0

#find how many pts above fitted line
for (i in 1:n0){
 if (y0[i]>y20[i]) {v0<-v0+1}
}

for (i in 1:n1){
 if (y1[i]>y21[i]) {v1<-v1+1}
}


#Calculate P-value
p<-choose(n0,v0)*choose(n1,v1)/choose(n,v0+v1)


output<-c(n0,n1,v0,v1,p)
print("n0, n1, v0, v1, p-value")
print(output)
```