ELSEVIER

Journal of Clinical Epidemiology 57 (2004) 461–469

Journal of Clinical Epidemiology

The conditional relative odds ratio provided less biased results for comparing diagnostic test accuracy in meta-analyses

Sadao Suzuki[a,b,*], Takeo Moro-oka[a], Niteesh K. Choudhry[c,d]

[a]*Department of Epidemiology, Harvard School of Public Heath, 677 Huntington Avenue, Boston, MA, 02115, USA*
[b]*Department of Health Promotion and Preventive Medicine, Nagoya City University Graduate School of Medical Sciences, Mizuho-ku, Nagoya 467-8601, Japan*
[c]*Department of Ambulatory Care and Prevention, Harvard Medical School, 133 Brookline Avenue, Boston, MA, 02215, USA*
[d]*Brigham and Women's Hospital, 75 Francis Street, Boston, MA, 02115, USA*

Accepted 22 September 2003

Abstract

Objective: Meta-analytic techniques are used to combine the results of different studies that have evaluated the accuracy of a given diagnostic test. The techniques commonly generate values that either describe the performance of a particular test or compare the discriminative ability of two tests. The later has received very little attention in the literature, and is the focus of this article.

Study Design and Setting: We summarize existing methods based on an odds ratio (OR) and propose a novel technique for conducting such analysis, the conditional relative odds ratio (CROR). We demonstrate how to extract the required data and calculate several different comparative indexes using a hypothetic example.

Results: A paired analysis is preferred to decrease selection bias and increase statistical power. There is no standard method of obtaining the standard error (SE) of each relative OR; thus, the SE of the summary index might be underestimated under the assumption of no within-study variability.

Conclusion: The CROR method estimates less biased indexes with SEs, and conditioned on discordant results, it is much less problematic ethically and economically. However, small cell counts may lead to larger SEs, and it might be impossible to construct McNemar's 2×2 tables for some studies.

Keywords: Meta-analysis; Diagnostic test; Sensitivity and specificity; Odds ratio; Paired comparison; ROC curve

1. Introduction

The most common way to describe the performance of a diagnostic test is the 2×2 table, which gives a number of positive and negative test results among the subjects with and without the disease. Diagnostic accuracy, which refers to the ability of a test to discriminate between subjects with and without disease, is commonly measured using sensitivity and specificity. These two parameters are negatively correlated, and vary according to the threshold value being considered. This relationship often is presented in the form of a receiver operating characteristic (ROC) curve, which plots sensitivity (the true positive rate, TPR) against $1 -$ specificity (the false positive rate, FPR) [1].

Meta-analytic techniques have been used for over a decade to combine the results of different studies that have all evaluated the accuracy of a given diagnostic test. The techniques commonly generate values that summarize the discriminative ability of the test, and that are calculated from a pair of TPR and FPR results for a given test threshold in the original studies. The values extracted from the original studies can be summarized in several ways, including using regression techniques to estimates a summary ROC (SROC) curve [1–3].

The summary statistics of diagnostic test accuracy generated in this way may be used either to describe the performance of a particular test or to compare the discriminative ability of two tests. Although the former has been well discussed [1–7], the later of these uses has received very little attention in the literature [3,8–10] and will, consequently, be the focus of this article.

2. Methods

An odds ratio (OR) provides an index of the discriminative ability of a test at a specified threshold. It is defined as the

* Corresponding author. Department of Health Promotion and Preventive Medicine, Nagoya City University Graduate School of Medical Sciences, Mizuho-ku, Nagoya 467-8601, Japan. Tel.: 81528538176; fax: 81528423830.

E-mail address: ssuzuki@med.nagoya-cu.ac.jp (S. Suzuki).

0895-4356/04/$ – see front matter
doi: 10.1016/j.jclinepi.2003.09.017

odds of test positivity among diseased subjects divided by odds of test positivity among nondiseased subjects, i.e.,

$$\begin{aligned}\mathrm{OR} &= \mathrm{odds(test(+)|diseased)/odds(test(+)|nondiseased)}\\ &= \mathrm{odds(true\ positive)/odds(false\ positive)}\end{aligned}$$

For mathematic convenience, the OR may be log-transformed such that $\log\mathrm{OR} = D$. When comparing two different diagnostic tests, the ratio of the ORs for the tests provides a measure of the relative accuracy of one test to the other. For the purposes of this article, our summary estimate will be a relative summary OR with a 95% confidence interval (CI). The weight for the summary OR in the fixed effects model and weighted regression will be assigned using the inverse variance of D. For each method, we provided a hypothetic example to show the actual steps of computation. The example of the test results is shown in Table 1. All analyses were conducted using SAS release 6.12 (SAS Institute, Cary, NC).

3. Approaches to comparison of two diagnostic tests

3.1. Comparison using the difference of two summary Ds (the traditional method of meta-analysis)

The traditional meta-analytic method for comparing two diagnostic tests is to perform a one-parameter analysis by extracting D values from each study and then summarizing Ds, taking into consideration the standard error (SE) of the individual D values. The summary D can be calculated using several weighting systems, such as an unweighted (equal weight) model, a fixed effects model [11–13], or a random effects model [13–15]. Two diagnostic tests can then be compared using the difference of summary Ds. These summary Ds and their difference are usually transformed back to the summary ORs or the relative ORs. The actual computational steps, with and without weighting, are shown in Table 2. The SAS program with several weighting is provided elsewhere [16].

3.2. Comparison using two summary Ds considering a threshold (the SROC method)

In the traditional method, we ignore the possibility that the OR varies with the test threshold. To take this into account, the use of an SROC curve has been proposed [1–3]. This method uses two parameters D and S, where $D = \mathrm{logit(TPR)} - \mathrm{logit(FPR)} = \log(\mathrm{OR})$, and $S = \mathrm{logit(TPR)} + \mathrm{logit(FPR)}$. S varies with the test threshold. By transforming (FPR, TPR) to (S, D), the SROC curve can be fit a straight line, $D = \alpha + \beta S$. A linear regression model, with or without weighting, can be used to estimate an intercept and a slope in the (S, D) space. The intercept (α) is an estimate for D when $S = 0$. If the null hypothesis for the slope (β) = 0 is defensible, D is considered to be constant over S, meaning the OR is constant regardless of threshold. In this case, we may remove S from our regression model, thereby making it a one-parameter analysis. If $\beta = 0$ is not defensible, we can evaluate the test performance by α which stands for the summary D when S is fixed to zero. Finally, we compare the summary Ds of both diagnostic tests using their difference. In our example, shown in Table 3, as $\beta = 0$ was defensible, we calculated a simple average D and the SE with and without weighting. Regression with no independent variable gives the same results by simply removing S from the model. In either case, the point estimate is the identical to that by the traditional method if the weight used is same.

3.3. Summary of the ratio of the test performance (the ROR method)

If the two tests to be compared were performed on "paired" subjects within each study, we can summarize test performance using a paired statistic. Because the OR represents the discriminative ability of the diagnostic tests, the relative odds ratio (ROR) is regarded as the relative accuracy of one test against the other. The ROR can be extracted from each study, and then log-transformed. The final result is obtained from the test of the null hypothesis that the summary $\log\mathrm{ROR} = 0$ using a paired t-test [3]. An unweighted model is usually used because the variance is not available. For this method, we generate a 2×2 table for each test in a given study and therefore do not consider the relationship between the test results of the individual patients. Using our example, the computational steps are shown in Table 4. Once again, the point estimate was identical to those obtained by the unweighted models of the traditional and SROC methods.

Table 1
A hypothetical study example of the results of Test X and Test Y among diseased and nondiseased subjects

	Test X				Test Y				Diseased	Nondiseased
Study	True positive (a)	False negative (b)	False positive (c)	True negative (d)	True positive (a')	False negative (b')	False positive (c')	True negative (d')	X(+)Y(−) (b)[a]	X(+)Y(−) (b')[a]
(1)	10	5	3	9	8	7	6	6	2	1
(2)	8	10	1	19	9	9	2	18	1	1
(3)	11	8	1	7	6	13	1	7	5	0
(4)	19	5	0	20	21	3	1	19	3	1
(5)	15	4	4	14	16	3	4	14	1	3
(6)	20	9	1	12	17	12	4	9	5	0
(7)	16	8	2	30	18	6	4	28	6	1

[a] Number of discordant subjects whose test result from Test X was positive and that from Test Y was negative among diseased (b) or nondiseased (b') group.

Table 2
Comparison by summary *D*s using the traditional method

• STEP 1: Extract 2 × 2 table counts from each study (add 0.5 to each cell count, if needed)

Test results	Diseased	Nondiseased
Positive	True positive (*a*)	False positive (*c*)
Negative	False negative (*b*)	True negative (*d*)

Study(1) Test X	Diseased	Nondiseased
Positive	10.5	3.5
Negative	5.5	9.5

• STEP 2: Calculate the quantities *D* and its SE for each study
$D = \log\text{OR} = \log(ad/bc) = \log(10.5 \cdot 9.5/3.5 \cdot 5.5) = 1.6452$
$\text{SE}(D) = \text{SE}(\log\text{OR}) = (a^{-1} + b^{-1} + c^{-1} + d^{-1})^{1/2} = (10.5^{-1} + 5.5^{-1} + 3.5^{-1} + 9.5^{-1})^{1/2} = 0.8173$

Study	$D_{\text{Test X}}$ (SE)	$D_{\text{Test Y}}$ (SE)
(1)	1.6452 (0.8173)	0.1252 (0.7474)
(2)	2.3536 (0.9648)	2.0015 (0.8152)
(3)	1.9117 (1.0023)	0.8786 (1.0139)
(4)	4.9792 (1.5106)	4.3802 (1.0248)
(5)	2.4068 (0.7602)	2.7207 (0.7984)
(6)	2.8894 (0.9491)	1.0837 (0.6816)
(7)	3.1647 (07817)	2.8918 (0.6821)

• STEP 3: Summarize the *D*s and SE, and transform back to ORs with 95% CIs

Model	$D_{\text{Test X}}$ (SE)	$\text{OR}_{\text{Test X}}$ (95%CI)	$D_{\text{Test Y}}$ (SE)	$\text{OR}_{\text{Test Y}}$ (95%CI)
Unweighted	2.764 (0.998)	15.86 (1.37–182.5)	2.011 (0.833)	7.47 (0.97–57.53)
Fixed effects	2.549 (0.343)	12.80 (5.52–29.66)	1.900 (0.300)	6.68 (3.20–13.94)
Random effects	2.537 (0.310)	12.64 (5.92–26.99)	1.964 (0515)	7.12 (2.02–25.13)

• STEP 4: Compare *D*s using the difference

Model	Difference of summary *D* (SE)	Ratio of the summary OR (95%CI)	*P*-value
Unweighted	0.752 (1.300)	2.12 (0.08–51.1)	.58
Fixed effects	0.649 (0.485)	1.91 (0.58–6.28)	.22
Random effects	0.572 (0.601)	1.77 (0.40–7.72)	.37

However, the *P*-value by the ROR method was much smaller than those by the other unweighted models above.

3.4. Summary of the conditional relative odds ratio (the CROR method)

If McNemar's 2 × 2 tables for both diseased and nondiseased subjects are available, we can calculate the relative test performance conditioned on counts of discordant test results. This method is typically applied to the studies that provide individual test results of both diagnostic tests. The McNemar's ORs for diseased and nondiseased among only discordant cells may be calculated as follows,

$$\text{OR}_{\text{Diseased}} = \frac{\text{odds}(\text{testX}(+)|\text{testY}(-), \text{diseased})/}{\text{odds}(\text{testY}(+)|\text{testX}(-), \text{diseased})}$$

$$\text{OR}_{\text{Nondiseased}} = \text{odds}(\text{testX}(+)|\text{testY}(-), \text{nondiseased})/\text{odds}(\text{testY}(+)|\text{testX}(-), \text{nondiseased})$$

We call the ratio of these two ORs the conditional relative odds ratio (CROR). The statistic represents the relative accuracy of test X against the test Y conditioning on the discordant subjects, because,

$$\text{CROR} = \text{OR}_{\text{Diseased}}/\text{OR}_{\text{Non-diseased}}$$

$$= \frac{\text{odds}(\text{testX}(+)|\text{testY}(-), \text{diseased})/\text{odds}(\text{testY}(+)|\text{testX}(-), \text{diseased})}{\text{odds}(\text{testX}(+)|\text{testY}(-), \text{nondiseased})/\text{odds}(\text{testY}(+)|\text{testX}(-), \text{nondiseased})}$$

$$= \frac{\text{odds}(\text{testX}(+)|\text{testY}(-), \text{diseased})/\text{odds}(\text{testX}(+)|\text{testY}(-), \text{nondiseased})}{\text{odds}(\text{testY}(+)|\text{testX}(-), \text{diseased})/\text{odds}(\text{testY}(+)|\text{testX}(-), \text{nondiseased})}$$

$$= \text{OR}(\text{testX}(+)|\text{testY}(-))/\text{OR}(\text{testY}(+)|\text{testX}(-))$$

= Relative accuracy of test X to test Y among discordant subjects.

The CROR from each study can be summarized in the same way as the traditional summary OR, with a specific weight. The CROR may differ from other relative ORs, because we extract different information from the original studies. In our examples, as shown in Table 5, we obtained a large point estimate and a different *P*-value than those obtained by the other methods. This difference is the same as

Table 3
Comparison by summary *D*s using the SROC method

• STEP 1: Extract 2 × 2 table counts from each study (add 0.5 to each cell count, if needed)

Test results	Diseased	Nondiseased
Positive	True positive (*a*)	False positive (*c*)
Negative	False negative (*b*)	True negative (*d*)

Study (1) Test X	Diseased	Nondiseased
Positive	10.5	3.5
Negative	5.5	9.5

• STEP 2: Calculate the quantities *D* with the SE and *S* for each study
$D = \log(a/b) - \log(c/d) = 1.6452$
$SE(D) = SE(\log OR) = (a^{-1} + b^{-1} + c^{-1} + d^{-1})^{1/2} = 0.8173$
$S = \log(a/b) + \log(c/d) = -0.3519$

Study	$D_{Test\ X}$ (SE)	$S_{Test\ X}$	$D_{Test\ Y}$ (SE)	$S_{Test\ Y}$
(1)	1.6452 (0.8173)	−0.3519	0.1252 (0.7474)	0.1251
(2)	2.3536 (0.9648)	−2.7762	2.0015 (0.8152)	−2.0014
(3)	1.9117 (1.0023)	−1.3071	0.8786 (1.0139)	−2.3403
(4)	4.9792 (1.5106)	−2.4479	4.3802 (1.0248)	−0.7496
(5)	2.4068 (0.7602)	0.0666	2.7207 (0.7984)	0.3805
(6)	2.8894 (0.9491)	−1.3511	1.0837 (0.6816)	−0.4107
(7)	3.1647 (0.7817)	−1.8381	2.8918 (0.6821)	−0.7998

• STEP 3: Fit a linear regression model and test $\beta = 0$
$D = \alpha + \beta S$ where α = intercept and β = slope

Regression	$\beta_{Test\ X}$ (SE)	*P*-value	$\beta_{Test\ Y}$ (SE)	*P*-value
Unweighted	−0.574 (0.403)	.21	0.097 (0.635)	.88
Weighted	−0.390 (0.311)	.26	−0.075 (0.636)	.91

If $\beta = 0$ is defensible, *S* may be removed from the model.

• STEP 4: Estimate summary *D* with/without weight and calculate the OR[a] summary $D = \alpha$

Regression	$D_{Test\ X}$ (SE)	$OR_{Test\ X}$ (95%CI)	$D_{Test\ Y}$ (SE)	$OR_{Test\ Y}$ (95%CI)
Unweighted	2.764 (0.418)	15.86 (5.69–44.19)	2.011 (0.547)	7.47 (1.95–28.54)
Weighted	2.549 (0.314)	12.80 (5.92–27.65)	1.900 (0.507)	6.68 (1.93–23.15)

• STEP 5: Compare *D*s using the difference

Regression	Difference of summary *D* (SE)	Ratio of the summary OR (95%CI)	*P*-value
Unweighted	0.752 (0.689)	2.12 (0.39–11.46)	.31
Weighted	0.649 (0.597)	1.91 (0.44–8.25)	.31

[a] In this example, *S* is removed from the model of step 4, because $\beta = 0$ is defensible.

that between the OR and McNemar's OR. If the ROR and the CROR differ, the CROR, like McNemar's OR, is adjusted for confounders know or unknown, and the CROR is methodologically unbiased. The SAS program for the CROR method using three weights is provided in the appendix.

3.5. Characteristic of each method

Each of the methods described above differs with respect to the data required for calculation and the underlying assumptions. In Table 6 we summarized the major characteristics of each method. As stated above, the ROR method and the unweighted model of the SROC method do not consider within-study variability, and the weighted model of the SROC method reflects just relative impact of original studies, although the traditional and SROC methods do not consider the correlation between two diagnostic tests.

4. Discussion

The validity of the comparison of two diagnostic tests by meta-analytic techniques clearly depends on the validity of the primary studies included in the analysis [17–20]. The study validity for diagnostic tests can be judged by considering the choice of gold standard, independence, and the likelihood of verification bias and selection bias [21,22]. Among these four criteria, we should be careful about selection bias when we apply unpaired tests. Because in unpaired tests, such as the traditional and SROC methods, the correlation between two tests is ignored, we can summarize studies even if they provided only results from either diagnostic test. This means that primary studies may be included in the analysis, even if they do not directly compare the two tests in question. Another threat to the validity of meta-analysis is publication bias, which has been reported to be a more serious problem in studies of diagnostic test accuracy

Table 4
Comparison using the summary relative odds ratio (ROR method)

• STEP 1: Extract a 2 × 2 table counts from each study (add 0.5 to each cell count, if needed)

Test X	Diseased	Nondiseased
Positive	True positive (a)	False positive (c)
Negative	False negative (b)	True negative (d)

Test Y	Diseased	Nondiseased
Positive	True positive (a')	False positive (c')
Negative	False negative (b')	True negative (d')

Test X	Diseased	Nondiseased
Positive	10.5	3.5
Negative	5.5	9.5

Test Y	Diseased	Nondiseased
Positive	8.5	6.5
Negative	7.5	6.5

• STEP 2: Calculate the OR, and log-transformed relative odds ratio (logROR)

$OR_{Test\ X} = ad/bc$

$OR_{Test\ Y} = a'd'/b'c'$

$logROR = \log(OR_{Test\ X}/OR_{Test\ Y}) = \log[(10.5 \cdot 9.5/5.5 \cdot 3.5)/(8.5 \cdot 6.5/7.5 \cdot 6.5)] = 1.5199$

Study	logROR
(1)	1.5199
(2)	0.3521
(3)	1.0331
(4)	0.5990
(5)	−0.3138
(6)	1.8057
(7)	0.2729

• STEP 3: Summarize logROR and Test summary logROR = 0 using the paired *t*-test

Model	Summary logROR (SE)	Summary ROR (95%CI)	*P*-value
Unweighted	0.752 (0.281)	2.12 (1.06–4.22)	.036

[22,23]. However, in comparison by meta-analysis, we can expect this bias to be less troublesome if the studies were paired because null results, that is, finding that two tests have the same accuracy, are as interesting as the non-null results. Therefore, negative comparison studies are more likely to be published than the results of single test studies. Of course, usual methods to detect publication bias, such as the funnel plot [24] or chronologic trend of effect [25] also can be used for these comparison studies.

How test thresholds are dealt with differs between those meta-analyses that describe and those that actually compare test performance. The concept of the SROC curve is to produce an ROC curve from several studies with varying thresholds. This allows for the calculation of a precise estimate, which is primarily useful for describing test performance. In contrast, to compare diagnostic tests we usually employ a fixed and commonly used threshold, and therefore we can construct a simpler and more pragmatic test hypothesis. If the threshold is fixed, S has a relatively narrow range, and D also tends to be constant. Thus, in many cases, $\beta = 0$ is defensible, and then S could be removed from the model. In this case, point estimate of D by the same weight is identical to that by the traditional method. If it is not defensible with a fixed threshold, there may be a factor affecting S and D. Consequently, a stratified analysis by this factor is recommended.

In the unweighted model, variability of the traditional model lies only within a study, while that of the SROC method lies only between studies. In the fixed effects model using the inverse variance weights, the variability of the traditional method still comes only from within a study, while that of the SROC method comes mainly from between studies and indirectly from within a study. The SE from the SROC method using a fixed effects weight just reflects relative stability of D from each study. For example, if all SEs from the original study are exactly doubled, the SE of summary D does not change. The random effects model takes both origins into account. If threshold is not fixed, it should be meaningful to test the null hypothesis for β before applying the traditional method. That is, if $\beta = 0$ is not defensible with a varying threshold, the SROC method is more appropriate, because the result is adjusted for the threshold. In that case, we usually compare Ds with fixed S to zero.

Unpaired tests summarize individual test accuracy from original studies first, and then compare the two values. In contrast, paired tests calculate the relative test accuracy within an original study first and then summarize these values. If the tests were applied to "paired" group(s) in all studies included in the meta-analysis, we can apply the ROR method which is more statistically powerful as long as D is positively correlated between studies. Empirically, the correlation is positive. Because the vagaries of random sampling of cases, which produce a higher/lower than expected accuracy index for one test, will also produce correspondingly higher/lower accuracy than one would expect for the second test [26]. Thus, we expect higher statistical

Table 5
Comparison using the summary conditional relative odds ratio (CROR method)

• STEP 1: Extract a McNemar's 2×2 tables by disease status (add 0.5 to each cell count, if needed)

	Test X			
	Diseased		Nondiseased	
Test Y	Positive	Negative	Positive	Negative
Positive	—	c	—	c'
Negative	b	—	b'	—

c = True positive$_{\text{Test Y}}$ − True positive$_{\text{Test X}}$ + b = 8 − 10 + 2 = 0[a]
c' = True negative$_{\text{Test X}}$ − True negative$_{\text{Test Y}}$ + b' = 9 − 6 + 1 = 4[a]

Study (1)	Test X			
	Diseased		Nondiseased	
Test Y	Positive	Negative	Positive	Negative
Positive	—	0.5	—	4.5
Negative	2.5	—	1.5	—

• STEP 2: Calculate the McNemar's ORs, and log-transformed conditional relative OR (logCROR) with the SE

$$\text{OR}_{\text{Diseased}} = b/c$$
$$\text{OR}_{\text{Nondiseased}} = b'/c'$$
$$\text{logCROR} = \log(\text{OR}_{\text{Diseased}}/\text{OR}_{\text{Nondiseased}}) = \log(2.5/0.5)/(1.5/4.5)] = 2.780$$
$$\text{SE(logCROR)} = (b^{-1} + c^{-1} + b'^{-1} + c'^{-1})^{1/2} = (2.5^{-1} + 0.5^{-1} + 1.5^{-1} + 4.5^{-1})^{1/2} = 1.8135$$

Study	logCROR (SE)
(1)	2.7080 (1.8135)
(2)	0.0000 (1.4605)
(3)	2.3979 (2.4863)
(4)	0.0588 (1.2386)
(5)	−0.5108 (1.2798)
(6)	2.7343 (1.6933)
(7)	0.5790 (1.1062)

• STEP 3: Summarize logCROR and test summary logCROR = 0

Model	Summary logCROR (SE)	Summary CROR (95%CI)	*P*-value
Unweighted	1.138 (0.536)	3.12 (0.83–11.6)	.078
Fixed effects	0.703 (0.545)	2.02 (0.53–7.67)	.24
Random effects	0.703 (0.545)	2.02 (0.53–7.67)	.24

[a] For the number of true positive and true negative for test X and test Y, see Table 1.

power with a narrower CI in the paired test. The ROR method, equivalent to the paired *t*-test, is easily performed and statistically powerful. However, we should be aware that this smaller *P*-value also depends on an implicit assumption of no within-study variability. If within-study variability is large, this method generates an incorrectly small SE. It is also important to note that we can technically calculate the ROR when two ORs are given in one study even if they are not "paired" or comparable, and therefore, we should be careful about the application of this method. For example, although it would be reasonable to apply the method to randomized controlled groups, or to the individuals on whom both tests were performed, it is not appropriate to apply this method to different individuals who were not assigned randomly.

To obtain the CROR, both diagnostic tests should be performed on paired individuals or, more practically, on the same individual. The CROR method has several advantages. First of all, we can condition out potential confounders because this method is stratified to the individual level using McNemar's ORs. Thus, we obtain a less biased estimate of relative test accuracy. Second, we can calculate the SE, which the ROR method does not provide, in a standard fashion. This means we can take within-study variability into account. Third, the CROR does not require the number of the concordant cells, and therefore, for comparison of two diagnostic tests, we do not need the true diagnosis if the results from two tests are concordant. This feature saves time and money for the follow-up of the test-negative subjects from both tests, which are usually much larger in number and clinically less important than discordant subjects. The fact that we do not need the true diagnosis for the individual having two concordant results means the CROR method can be performed ethically. For example, when we compare two screening tests, the method clears the potential ethical problem to perform further tests to the test negative subjects. At the same time, it saves much cost, which is one of the biggest advantages of the CROR method. However, precisely because this test is based on only the discordant individuals, the cell count tends to be small, and thus the SE may be larger. Another weakness of this method is that it might be impossible to construct McNemar's 2×2 tables in several studies, which may significantly limit the number of studies to be included in the meta-analysis.

In the SROC method, an unweighted model is recommended for two reasons. First, it gives similar results to the random effects model in the traditional method [3]. This recommendation relies on an assumption that between-study variability is much larger than within-study variability. If this assumption is not defensible, the SE in the unweighted model in the SROC method is incorrectly small. It is important to understand the SE from the SROC method has a totally different origin than the traditional method and thus, may generate different results. Second, weighting by the inverse variance may bias the estimate [7]. That is, at equivalent sample sizes, studies that appear to show poorer accuracy will be given more weight. Consequently, the ORs calculated by the weighted model are often lower than those from the unweighted model. This also occurs in the fixed effects and random effects models in the traditional method. In paired analysis, we do not have this problem because we compare the accuracy within a study before summarizing across studies. However, in the weighted analysis of the CROR method, studies with a CROR of approximately 1

Table 6
Required data and characteristics of each method by weight

	Extracted table(s) from an original study		
Method weight	Data needed	Variability taken into account	Correlation of two test results
Traditional method	2×2 tables of both tests		Not considered
Unweighted	Ds with SEs of both tests	Within-study variability only	(independent)
Fixed effects	Ds with SEs of both tests	Within-study variability only	
Random effects	Ds with SEs of both tests	Within-and between-study variability	
SROC method	2×2 table of both tests		Not considered
Unweighted	Ds and Ss of both tests	Between-study variability only	(independent)
Weighted	Ds with SEs and Ss of both tests	Mainly between-study variability	
ROR method	2×2 table of both tests		Group level
Unweighted	logROR (difference of Ds)	Between-study variability only	
CROR method	2×2 McNemar's table by disease status		Individual level
Unweighted	logCROR with SE	Within-study variability only	
Fixed effects	logCROR with SE	Within-study variability only	
Random effects	logCROR with SE	Within-and between-study variability	

are often relatively heavily weighted for the number of discordant individuals, and thus, an unweighted model usually provides the most powerful results. Therefore, despite the disadvantage of not paying more attention to larger studies, unweighted model seems to be more appropriate. Methods of weighting which avoid this problem need to be explored.

When summarizing the OR, we extract 2×2 tables from the original studies and then perform our analytic procedures. If we were to create a pooled 2×2 table by adding up all the subjects by cells, this may generate bias by confounding even if the individual discriminative power (OR) is same [22]. This is identical to the confounding of the crude ORs served in etiologic studies, which can be avoided by summarizing the OR. Confounding may also be generated in the CROR method when the results were pooled. In our example, the CROR from study (1) and study (3) were 15 and 11, respectively, and the pooled CROR was 22.5. Heterogeneity lies not only between studies but also within a study. If within-study heterogeneity generates bias, we should stratify the subjects by the confounder, and summarize separated ORs.

It is important to note that the ability of diagnostic tests to rule in or rule out disease, usually captured by the sensitivity and the specificity, are also captured in the course of the CROR method. The CROR is the ratio of the positive odds ratio among diseased subjects ($OR_{Diseased}$) to that among nondiseased ($OR_{Nondiseased}$). The $OR_{Diseased}$ means the relative odds of the true positive results of the two diagnostic tests, while the $OR_{Nondiseased}$, is the relative odds of the false positive results of them. Therefore, we can obtain information of the asymmetric nature of the tests using CROR method as well as the other methods, which provide the sensitivity and the specificity from 2×2 tables. For example, a new test may be good in ruling in a diagnosis and poor in excluding the diagnosis, while the opposite may be the case for the standard test. In this case, the $OR_{Diseased}$ is larger than 1, as is the $OR_{Nondiseased}$. The purposes of the two kinds of index, for the total diagnostic ability or ability for ruling in/out disease, are not the same, and we should be careful to the application of them.

Another meta-analytic approach not describe above is to summarize the area under a curve (AUC) for the ROC curve from each study [27–29]. If two tests were performed on the same individuals, we can calculate the difference of the AUC with its SE [26] and summarize the difference afterwards. This method integrates out the threshold and evaluates test performance as a whole. However, the AUC might be influenced by the extreme thresholds that are not used in practical diagnosis, so we should be careful about the shape of the original ROC curve. This method has another limitation that there are few papers that provide the AUC with the SE, or enough information to obtain these estimates. The generalized estimating equation [30] may be useful when data are available for two or more thresholds. Recently, estimating a SROC curve using a bivariate random effects model [31] has been proposed, which takes α and β into account simultaneously.

In summary, in this article, we have reviewed and made recommendations regarding the comparison of two diagnostic tests by meta-analytic technique based on the OR or the relative OR. We proposed a novel test, the CROR, which is statistically less biased and takes into account within-study variability. Limitations of this method are the wider CI due to the smaller numbers of discordant subjects and more information required from the original studies, which may decrease the number of studies included in the meta-analysis.

Acknowledgments

The authors thank Drs. Sophie Michaud and Stephan Harbarth for providing an opportunity to think about the CROR method and Dr. Nan Laird for the helpful comments on the OR and the relative OR from statistical view points.

Appendix

```
data data1;
 input id tpx fnx fpx tnx tpy fny fpy tny xnypd xnypnd;
 xpynd=tpy-tpx+xnypd; xpynnd=tnx-tny+xnypnd;
 d=log((xnypd+0.5)/(xpynd+0.5)*(xpynnd+0.5)/(xnypnd+0.5));
 dse=sqrt(1/(xnypd+0.5)+1/(xpynd+0.5)+1/(xpynnd+0.5)+1/(xnypnd+0.5));
 w=1/(dse**2); ww=w**2; wd=w*d;
cards;
1 10  5 3  9  8  7 6  6 2 1
2  8 10 1 19  9  9 2 18 1 1
3 11  8 1  7  6 13 1  7 5 0
4 19  5 0 20 21  3 1 19 3 1
5 15  4 4 14 16  3 4 14 1 3
6 20  9 1 12 17 12 4  9 5 0
7 16  8 2 30 18  6 4 28 6 1
;
run;

title 'Conditional Relative Odds Ratio (CROR)';
title2 'Unweighted Model';
proc means noprint mean stderr t prt; var d;
output out=out mean=slcroru stderr=seu n=n; run;

data data2; set out;
 t=tinv(0.975, n-1); scror=exp(slcroru);
 lowscror=exp(slcroru-t*seu); uppscror=exp(slcroru+t*seu);
 pu=2*(1-probt((slcroru/seu), n-1));
 proc print; var slcroru seu scror lowscror uppscror pu; run;

title2 'Fixed Effects Model';
proc means data=data1 noprint sum n mean noprint; var wd w ww;
 output out=out1 sum=swd sw sww n=n; run;

data data2; set data1; if _n_=1 then set out1;
 slcrorf=swd/sw; sef=sqrt(1/sw); scrorf=exp(slcrorf); t=tinv(0.975, n-1);
 loscrorf=exp(slcrorf-t*sef); upscrorf=exp(slcrorf+t*sef);
 pf=2*(1-probt((slcrorf/sef), n-1)); q=w*(d-slcrorf)**2; run;

proc print; where id=1; var slcrorf sef scrorf loscrorf upscrorf pf; run;

title2 'Random Effects Model';
proc means sum noprint; var q; output out=out2 sum=sq; run;

data data3; set data2; if _n_=1 then set out2; n=_freq_;
 h=max(0,(sq-(n-1))*sw/((sw**2)-sww)); wr=1/(h+(1/w)); num=wr*d; run;

proc means sum noprint; var num wr;
 output out=out3 sum=snum swr n=n; run;

data data4; set out3;
 slcrorr=snum/swr; ser=sqrt(1/swr);
 scrorr=exp(slcrorr); t=tinv(0.975, n-1);
 loscrorr=exp(slcrorr-t*ser); upscrorr=exp(slcrorr+t*ser);
 pr=2*(1-probt((slcrorr/ser), n-1)); run;

proc print; var slcrorr ser scrorr loscrorr upscrorr pr; run;
```

References

[1] Littenberg B, Moses LE. Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method. Med Decis Making 1993;13:313–21.

[2] Kardaun JWPF, Kardaun OJWF. Comparative diagnostic performance of three radiological procedures for the detection of lumber disk herniation. Method Inform Med 1990;29:12–22.

[3] Moses LE, Shapiro D, Littenberg B. Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. Stat Med 1993;12:1293–316.

[4] Vamvakas EC. Meta-analysis of studies of the diagnostic accuracy of laboratory tests. Arch Pathol Lab Med 1988;122:675–86.

[5] Shapiro DE. Issues in combining independent estimates of the sensitivity and specificity of a diagnostic test. Acad Radiol 1995;2:S37–S47.

[6] Irwig L, Tosteson AN, Gatsonis C, Lau J, Colditz G, Chalmers TC, Mosteller F. Guidelines for meta-analyses evaluating diagnostic tests. Ann Intern Med 1994;120:667–76.

[7] Irwig L, Macaskill P, Glasziou P, Fahey M. Meta-analytic methods for diagnostic test accuracy. J Clin Epidemiol 1995;48:119–30.

[8] Scouller K, Conigrave KM, Macaskill P, Irvig L, Whitfield JB. Should we use carbohydrate-deficient transferrin instead of γ-glutamyltransferase for detecting problem drinkers? A systematic review and meta-analysis. Clin Chem 2000;46:1894–902.

[9] Hallan S, Åsberg A. The accuracy of C-reactive protein in diagnosing acute appendicitis—a meta-analysis. Scand J Clin Lab Invest 1997;57:373–80.

[10] Dwamena BA, Sonnad SS, Angobaldo JO, Wahl RL. Metastases from non-small cell lung cancer: mediastinal staging in the 1990s—meta-analytic comparison of PET and CT. Radiology 1999;213:530–6.

[11] Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Cancer Inst 1959;22:719–48.

[12] Yusuf S, Peto R, Lewis J, Collins R, Sleight P. Beta blockade during and after myocardial infarction: an overview of the randomized trials. Prog Cardiovasc Dis 1985;27:335–71.

[13] Greenland S. Meta-analysis. In: Rothman KJ, Greenland S, editors. Modern epidemiology. 2nd ed. Philadelphia: Lippincot-Raven; 1998. p. 643–709.

[14] DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials 1986;7:177–88.

[15] Fleiss JL, Gross AJ. Meta-analysis in epidemiology, with special reference to studies of association between exposure to environmental tobacco smoke and lung cancer: A critique. J Clin Epidemiol 1991;44:127–39.

[16] Shadish WR, Haddock CK. Combining estimates of effect size. In: Cooper H, Hedges AV, editors. The handbook of research synthesis. New York: Russell Sage Foundation; 1994. p. 261–81.

[17] Sheps SB, Schecher MT. The assessment of diagnostic tests. JAMA 1984;252:2418–22.

[18] Arroll B, Schecher MT, Sheps SB. The assessment of diagnostic tests: a comparison of the recent medical literature—1982 versus 1985. J Gen Intern Med 1988;3:443–7.

[19] Cooper LS, Chalmers TC, McCally M, Berrier J, Sacks HS. The poor quality of early evaluations of magnetic resonace imaging. JAMA 1988;259:3277–80.

[20] Beam CA, Sosman HD, Zheng JY. Status of clinical MRI evaluations, 1985–1988: baseline and design for further assessments. Radiology 1991;180:265–9.

[21] Sackett DL, Haynes RB, Guyatt GH, Tugwell P. Clinical epidemiology: a basic science for clinical medicine. 2nd ed. Boston, MA: Little Brown; 1991.

[22] Begg CB. Biases in the assessment of diagnostic tests. Stat Med 1987;6:411–23.

[23] Begg CB, Berlin JA. Publication bias and dissemination of clinical research. J Natl Cancer Inst 1989;81:107–15.

[24] Light RJ, Pillemer DB. Summing-up. The science of reviewing research. Cambridge: Harvard University Press; 1984.

[25] Jenicek M. Meta-analysis in medicine, putting experience together. In: Epidemiology, the logic of modern medicine. Montreal: Epimed International; 1995. p. 269–95.

[26] Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983;148:839–43.

[27] Bamber D. The area above the ordinal dominance graph and the area below the receiver operating graph. J Math Psychol 1975;12:387–415.

[28] McClish DK. Combining and comparing area estimates across studies or strata. Med Decis Making 1992;12:274–9.

[29] Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29–36.

[30] Zeger S, Liang KY. Longitudinal data analysis for discrete and continuous outcomes. Biometrics 1986;42:121–30.

[31] Kester ADM, Buntinx F. Meta-analysis of ROC curves. Med Decis Making 2000;20:430–9.