

Meta-analyses of diagnostic studies

Ton J. Cleophas^{1,2,*} and Aeilko H. Zwinderman^{1,3}

¹ European Interuniversity College of Pharmaceutical Medicine, Lyon, France

² Albert Schweitzer Hospital Dordrecht, Dordrecht, The Netherlands

³ Academic Medical Center, Amsterdam, The Netherlands

Abstract

Background: Diagnostic reviews often include the sensitivity/specificity results of individual studies. A problem occurs when these data are pooled because the correlation between sensitivity and specificity is generally strongly negative, causing overestimation of the pooled results. The diagnostic odds ratio (DOR), defined as the odds of true positives vs. that of false positives, may avoid this problem. The aim of the study was to review the advantages and limitations of the DORs.

Methods: A systematic review of 44 previously published diagnostic studies was used as an example.

Results: DORs can be readily implemented in diagnostic research. Advantages include: (1) they adjust for the negative and curvilinear correlations between sensitivities and specificities, (2) they take account of the heterogeneity between studies with respect to the different thresholds chosen by the investigators in the original studies, and (3) it is easy to extend the model with covariates representing between-study differences in design. Limitations include: 1) the outcome parameter is a summary estimate of both sensitivity and specificity, and 2) the magnitude of the studies included is not taken into account.

Conclusions: Reported sensitivities and specificities of different studies assessing similar diagnostic tests are not only negatively correlated, but also negatively correlated in a curvilinear manner. It is appropriate to take this negative curvilinear correlation into account in the data pooling of such meta-analyses. The DORs can be applied for that purpose.

Clin Chem Lab Med 2009;47:1351–4.

Keywords: diagnostic odds ratios; false negative rate; false positive rate; meta-analysis; qualitative diagnostic studies; ROC curves; sensitivity; specificity; summary ROC curves; true negative rate; true positive rate.

*Corresponding author: Ton J. Cleophas, MD, PhD, Professor, European College of Pharmaceutical Medicine, Lyon France, c/o Department Medicine Albert Schweitzer Hospital, Dordrecht Netherlands, Box 444, 3300 AK Dordrecht, Netherlands
Phone: +31 184 43422, Fax: +31 184 434340, E-mail: ajm.cleophas@wxs.nl
Received June 1, 2009; accepted July 24, 2009; previously published online October 12, 2009

Introduction

In the past few years many novel diagnostic methods have been developed, including multi-slice computer tomography, magnetic resonance, positive emission tomography and many others. Studies evaluating their respective sensitivities and specificities have been published, and meta-analyses of these studies can be performed to establish whether the findings are consistent and can be generalized across populations and variations in morbidity/treatment. Sensitivity and specificity are estimators of accuracy of diagnostic methods, as explained in the following diagram.

	Gold standard test	Positive	Negative
Diagnostic test	Positive	TP	FP
	Negative	FN	TN

TP, number of true positive; FP, false positive; FN, false negative; and TN, true negative patients in a study.

Sensitivity = true positive rate (TPR) = $TP / (TP + FN)$
Specificity = true negative rate (TNR) = $TN / (TN + FP)$.

1-Specificity = $(TN + FP) / (TN + FP) - TN / (TN + FP)$
= $FP / (TN + FP) = FPR$

An intuitive approach to meta-analysis of diagnostic studies is to pool the odds of sensitivity (=TPR/(1-TPR)) and specificity (=TNR/(1-TNR)) of the separate studies. Sensitivities and specificities are, however, dependent on one another. In addition, they are dependent in a non-linear manner as shown in the summary receiver operated characteristic (ROC) curve in Figure 1. In order to account for these problems Moses and Wittemberg proposed diagnostic odds ratios (DORs) of the sensitivities vs. specificities (1). In recent years, this approach has been increasingly used (2–6). The current paper uses the results of a previously published review of diagnostic studies (7) as an example, and reviews the advantages and disadvantages of this novel method and discusses alternative possibilities.

Diagnostic odds ratios

The accuracy of a diagnostic test is usually summarized by two statistics: TPR or sensitivity, and the TNR or specificity. They are often used to draw ROC curves (Figure 1). Instead of the dual approach of sensitivity and specificity, accuracy can also be summarized by the DOR:

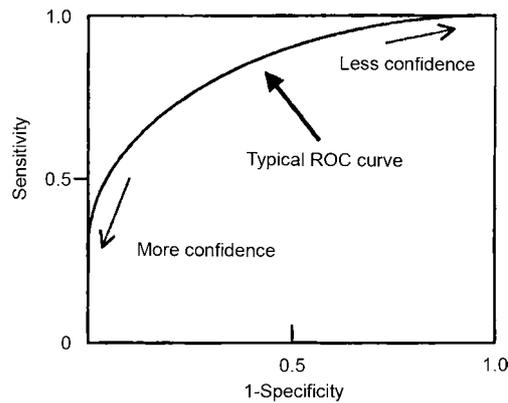


Figure 1 Example of a summary receiver-operated-characteristic (ROC) curve.

The proportion of true positive patients (=sensitivity) is drawn against the proportion of false positive patients (=1-specificity) using the results of multiple studies. With many diagnostic tests, test results do not necessarily fall into one of two categories, but rather into categories with variable confidence in the presence of a disease.

$$DOR = \frac{\text{sensitivity}/(1-\text{sensitivity})}{(1-\text{specificity})/\text{specificity}}$$

The DOR is an interesting term. It compares the odds of true positive patients with that of false positives, and, thus summarizes the overall accuracy of a diagnostic test. A problem is that, like any odds ratio, it does not follow a Gaussian distribution and logarithmic transformation is required. Linear regression of the $\ln(DOR)$ on the statistic S is often applied for analysis.

$$\ln(DOR) = \ln\left(\frac{TPR}{1-TPR}\right) - \ln\left(\frac{FPR}{1-FPR}\right) \text{ and}$$

$$S = \ln\left(\frac{TPR}{1-TPR}\right) + \ln\left(\frac{FPR}{1-FPR}\right)$$

where TPR and FPR are the true and false positive rates, and \ln refers to the natural logarithm. Linear regression analysis is simply fitting a straight line: $\ln(DOR) = a + b.S$. $\ln(DOR)$ is the dependent and S the independent variable. Although this is not obvious from the model as given, this method is often successful in producing a rather close linear fit to the data. The a -value is the intercept and the b -value is the regression coefficient. If sensitivity equals specificity, then $TPR = TNR = 1 - FPR$, and S reduces to 0. Thus, the magnitude of the $\ln(DOR)$ at that point equals the a -value. The DOR can be calculated by back-log-transformation of the calculated intercept.

As an example, results from a previously published review of 44 diagnostic studies (7) of imaging techniques for lymph node metastases are used (Tables 1-3). The $\ln(DOR)$ and S -values calculated from the lymphangiography-studies are entered into the SPSS software program. We command statistics; regression; linear. The program produces an a -value of 2.09 [standard error (SE)=0.35]. The DOR at the point where $S=0$ is then found by taking the invert of the

Table 1 Example of meta-analysis of 17 diagnostic studies of lymphangiography for assessment of lymph node metastases.

Study no.	TP	FP	FN	TN
1	0	1	6	17
2	12	3	3	7
3	4	1	2	13
4	10	4	3	25
5	3	1	4	12
6	9	3	3	29
7	20	4	8	31
8	17	5	7	21
9	2	0	9	32
10	3	1	9	38
11	1	1	2	18
12	5	2	2	61
13	21	8	40	184
14	4	3	9	42
15	0	0	5	15
16	7	11	22	158
17	3	3	2	29

TP, true positive; FP, false positive; FN, false negative; TN, true negative.

Table 2 Example of meta-analysis of 17 diagnostic studies of computerized tomography (CT) imaging of lymph nodes metastases.

Study no.	TP	FP	FN	TN
1	19	1	10	81
2	8	9	2	13
3	41	1	12	49
4	5	1	2	18
5	45	58	32	165
6	8	6	2	32
7	5	8	1	7
8	15	17	11	52
9	16	11	8	24
10	4	8	2	25
11	4	12	10	70
12	10	4	4	55
13	2	5	6	23
14	7	10	7	30
15	4	50	12	135
16	8	3	1	37
17	4	3	0	14

TP, true positive; FP, false positive; FN, false negative; TN, true negative.

Table 3 Example of meta-analysis of 10 diagnostic studies of magnetic resonance imaging (MRI) of lymph node metastases.

Study no.	TP	FP	FN	TN
1	9	2	2	41
2	3	6	5	32
3	3	2	1	16
4	3	1	12	44
5	0	0	5	15
6	7	2	22	167
7	12	4	4	29
8	23	5	14	230
9	8	5	5	53
10	16	2	2	22

TP, true positive; FP, false positive; FN, false negative; TN, true negative.

Table 4 Intercepts (a-value) and slopes (b-value) of the linear regression lines of the DORs of the three diagnostic modalities from Tables 1 through 3.

Diagnostic modality	Intercept (SE) (a-value)	Regression (SE) coefficient (b-value)	DOR at S=0 (SE) (p-value)
Lymphangiography	2.09 (0.30)	-0.35 (0.20)	8.08 (1.35) (<0.001 vs. CT and MRI)
CT	2.84 (0.44)	0.23 (0.14)	17.16 (1.55) (<0.001 vs. MRI and lymphangiography)
MRI	3.51 (0.56)	0.25 (0.17)	33.45 (1.75) (<0.001 vs. CT and lymphangiography)

DOR, diagnostic odds ratio; SE, standard error; CT, computed tomography; MRI, magnetic resonance imaging.

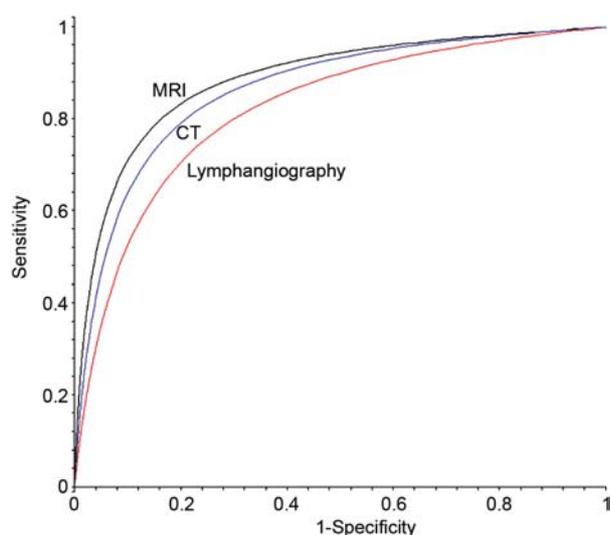
natural logarithm of $2.09=8.08$ (SE=1.35). A summary of the results of the regression analyses are shown in Table 4. The magnitude of DORs at $S=0$ can be used to estimate the degree of overall accuracy of the diagnostic method. Table 4 shows that magnetic resonance imaging (MRI) is significantly more accurate than the other two methods of cardiac imaging with $p<0.001$.

Constructing summary ROC curves

The results of the separate studies are used to calculate the best fit a and b for the data. Subsequently, the equation shown below is adequate to construct the best-fit summary ROC curve from the a - and b -values:

$$TPR = \left[1 + e^{-a(1-b)} \left(\frac{1-FPR}{FPR} \right)^{(1+b)(1-b)} \right]^{-1}$$

The equation is entered into Maple 9.5 software program for making graphs and we indicate the a - and b -value. The software program produces the best

**Figure 2** Summary ROC curves for the three diagnostic modalities.

A diagonal line drawn from the top of the y -axis to the right end of the x -axis contains the points of summary ROC curves where sensitivity equals specificity, and thus $S=0$. Along this diagonal line, the distance of the MRI curve to the top of the y -axis is shorter than that of the other curves, indicating a better accuracy of this diagnostic method. This is supported by a significantly larger DOR at $p<0.001$ (Table 4).

fit ROC curves for the three diagnostic methods (Figure 2). The curve closest to the top of the y -axis provides the best overall accuracy. A diagonal line from the top of the y -axis to the right end of the x -axis contains all the points on the summary ROC curves where sensitivity equals specificity, and thus $S=0$. Along this diagonal line, the distance from the MRI curve would be shorter than that of the other curves, indicating better accuracy of this diagnostic method. This is supported by a significantly larger DOR at $p<0.001$ as shown in Table 4, and discussed in the section above. The distances from the top of the y -axis to the MR/CT/lymphangiography summary ROC curves can be calculated using Pythagoras' equation for rectangular triangles, where

$$\sqrt{[(1-\text{sensitivity})^2 + (1-\text{specificity})^2]}, \text{ and equals:}$$

for the MR curve $\sqrt{(0.18^2 + 0.18^2)} = 0.25$,
for the CT curve $\sqrt{(0.22^2 + 0.22^2)} = 0.31$,
for the lymphangiography curve $\sqrt{(0.26^2 + 0.26^2)} = 0.37$.

Results and discussion

This paper shows that DORs can be readily implemented in the meta-analyses of diagnostic studies. An advantage of the DOR approach is that it incorporates the special correlation between sensitivities and specificities of the included studies. Another advantage is that it takes into account the heterogeneity between studies with respect to the different thresholds chosen by the investigators in the original studies. For particular reasons, some investigators prefer high sensitivity and accept low specificity, while others prefer the reverse. These differences produce heterogeneity in the DOR between studies, but this is taken into consideration by regression of $\ln(\text{DOR})$ on S . A subsequent advantage is that it is easy to extend the model with covariates representing differences in design between studies. However, this advantage is limited because in the summary ROC model, these covariates are supposed to affect sensitivity and specificity in a similar manner, which is not always the case.

We need to mention some limitations. First, since the outcome parameter is a summary estimate of both sensitivity and specificity, no summary estimates of sensitivity or specificity are available. Second, the magnitude of the studies included in the meta-analyses is not taken into account in the summary ROC method. Thus, it is impossible to weigh the true positives and false positives of the studies separately (8). However, it is common to report sensitivity

and specificity from diagnostic studies without accounting for the size of the sample from which they were calculated (9).

Other commonly used test indicators for diagnostic tests include positive predictive values, likelihood ratios and Youden's index. They are, in theory, useful for diagnostic meta-analyses, particularly from a bayesian perspective. However, they are rarely used because they are not practical from a statistical viewpoint due to numerical and ceiling problems (10).

As an alternative, multivariate methods for pooling the meta-data accounting sensitivities and specificities can be used. For example, multivariate methods such as multivariate analysis of variance (MANOVA) with sensitivity and specificity as outcome variables and different diagnostic modalities as predictor variable can produce results similar to those of the DOR method. In addition, it produces sensitivities and specificities separately and adjusted for their interaction (8). Again, a limitation with this approach is that the magnitude of the separate studies is not accounted for, and the numbers of studies included in the meta-analyses is often too small for reliable testing. A rule of thumb, at least 10 studies per variable are required for multivariate analyses.

We should add that, similar to therapeutic meta-analyses, it is appropriate to apply scientific rules in any meta-analysis of diagnostic studies including a thorough search of the literature, strict inclusion criteria and an assessment of the usual pitfalls of meta-analyses such as publication bias, clinical heterogeneity, and lack of robustness (11).

Conclusions

Reported sensitivities and specificities of various studies that assess similar diagnostic tests are not only negatively correlated, but also correlated in a

curvilinear manner. It is appropriate to take this negative curvilinear correlation into account in the pooling of data for meta-analyses. DORs can be applied for this purpose.

References

1. 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.
2. Hasselblad V, Hedges LV. Meta-analysis of screening and diagnostic tests. *Psychol Bull* 1995;117:167-78.
3. Irwig L, Macaskill P, Glasziou P, Fahey M. Meta-analytic methods for diagnostic test accuracy. *J Clin Epidemiol* 1995;48:119-30.
4. Walter SD. Properties of the summary receiver operating characteristic (SROC) curve for diagnostic test data. *Stat Med* 2002;21:1237-56.
5. Glas AS, Lijmer JG, Prins MH, Bossuyt PM. The diagnostic odds ratio: a single indicator of test performance. *J Clin Epidemiol* 2003;56:1129-35.
6. Bipat S, Glas AS, Van der Velden J, Zwinderman AH, Bossuyt PM, Stoker J. Computed tomography and magnetic resonance imaging in staging of uterine cervical carcinoma: systematic review. *Gynecol Oncol* 2003;91:59-66.
7. Scheidler J, Hricak H, Yu KK, Subak L, Segal MR. Radiological evaluation of lymph node metastases in patients with cervical cancer. A meta-analysis. *J Am Med Assoc* 1997;278:1096-101.
8. Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *J Clin Epidemiol* 2005;58:982-90.
9. Levin MD, Van de Bos E, Van Ouwerkerk BM, Cleophas TJ. Uncertainty of diagnostic tests. *Perfusion* 2008;21:42-8.
10. Cleophas TJ, Zwinderman AH, Van Ouwerkerk BM. Log likelihood ratio tests for the assessment of cardiovascular events. *Perfusion* 2007;20:79-82.
11. Cleophas TJ, Zwinderman AH. Primer in statistics: meta-analysis. *Circulation* 2007;115:2870-5.