ELSEVIER

Journal of Clinical Epidemiology 56 (2003) 1129–1135

Journal of Clinical Epidemiology

The diagnostic odds ratio: a single indicator of test performance

Afina S. Glas[a,*], Jeroen G. Lijmer[b], Martin H. Prins[c], Gouke J. Bonsel[d], Patrick M.M. Bossuyt[a]

[a]*Department of Clinical Epidemiology & Biostatistics, University of Amsterdam, Academic Medical Center, Post Office Box 22700, 100 DE Amsterdam, The Netherlands*

[b]*Department of Psychiatry, University Medical Center, Post Office Box 85500, 3508 GA, Utrecht, The Netherlands*

[c]*Department of Epidemiology, University of Maastricht, Post Office Box 6166200 MD, Maastricht, The Netherlands*

[d]*Department of Public Health, Academic Medical Center, Post Office Box 22700, 1100 DE, Amsterdam, The Netherlands*

Accepted 17 April 2003

Abstract

Diagnostic testing can be used to discriminate subjects with a target disorder from subjects without it. Several indicators of diagnostic performance have been proposed, such as sensitivity and specificity. Using paired indicators can be a disadvantage in comparing the performance of competing tests, especially if one test does not outperform the other on both indicators. Here we propose the use of the odds ratio as a single indicator of diagnostic performance. The diagnostic odds ratio is closely linked to existing indicators, it facilitates formal meta-analysis of studies on diagnostic test performance, and it is derived from logistic models, which allow for the inclusion of additional variables to correct for heterogeneity. A disadvantage is the impossibility of weighing the true positive and false positive rate separately. In this article the application of the diagnostic odds ratio in test evaluation is illustrated. © 2003 Elsevier Inc. All rights reserved.

Keywords: Diagnostic odds ratio; Tutorial; Diagnostic test; Sensitivity and specificity; Logistic regression; Meta-analysis

1. Introduction

In an era of evidence-based medicine, decision makers need high-quality data to support decisions about whether or not to use a diagnostic test in a specific clinical situation and, if so, which test. Many quantitative indicators of test performance have been introduced, comprising sensitivity and specificity, predictive values, chance-corrected measures of agreement, likelihood ratios, area under the receiver operating characteristic curve, and many more. All are quantitative indicators of the test's ability to discriminate patients with the target condition (usually the disease of interest) from those without it, resulting from a comparison of the test's results with those from the reference standard in a series of representative patients. In most applications, the reference standard is the best available method to decide on the presence or absence of the target condition. Less well known is the odds ratio as a single indicator of test performance. The odds ratio is a familiar statistic in epidemiology, expressing the strength of association between exposure and disease. As such it also can be applied to express the strength of the association between test result and disease.

This article offers an introduction to the understanding and use of the odds ratio in diagnostic applications. In brief, we will refer to it as the diagnostic odds ratio (DOR). First, we will point out the usefulness of the odds ratio in dichotomous and polychotomous tests. We will then discuss the use of the DOR in meta-analysis and the application of conditional logistic regression techniques to enhance the information resulting from such analysis.

2. Dichotomous test outcomes

Although most diagnostic tests have multiple or continuous outcomes, either grouping of categories or application of a cutoff value is frequently applied to classify results into positive or negative. Such a dichotomization enables one to represent the comparison between a diagnostic test and its reference standard in one 2 × 2 contingency table, as depicted in Table 1. Common indicators of test performance

* Corresponding author. Tel.: +31-(0)-20-566958; fax: +31-(0)-20-6912683.

E-mail address: a.s.glas@amc.uva.nl (A.S. Glas).

0895-4356/03/$ – see front matter © 2003 Elsevier Inc. All rights reserved.
doi: 10.1016/S0895-4356(03)00177-X

Table 1
2×2 contingency table

		Reference test	
		Target disorder	No target disorder
Test	positive	TP	FP
	negative	FN	TN

The abbreviations TP, FP, FN, and TN denote the number of respectively, true positives, false positives, false negatives, and true negatives. The same definitions are used throughout the text and Table 2.

derived from such a 2×2 table are the sensitivity of the test, its specificity, the positive and negative predictive values, and the positive and negative likelihood ratios [1]. (See Table 2 for a definition of these indicators.)

Unfortunately, none of these indicators in itself validly represent the test's discriminatory performance. Sensitivity is only part of the discriminatory evidence, as high sensitivity may be accompanied by low specificity. Additionally, no simple aggregation rule exists to combine sensitivity and specificity into one measure of performance.

Table 3 shows the performance of three different radiologic diagnostic tests to stage ovarian cancer as an illustration of the need for combined judgment. All three tests were performed in a group of 280 patients suspected of ovarian cancer [2]. Surgical and histopathologic findings were used as the reference standard. The sensitivity of the ultrasound was worse than that of the computer tomography (CT) scan in detecting peritoneal metastases, but for the specificity, the reverse held. Likelihood ratios and the predictive values are also not decisive. Also, the combined evidence of the pairs of indicators cannot simply be ranked.

For this, a single indicator of test performance like the test's accuracy is required. In addition to its global meaning of agreement between test and reference standard, accuracy in its specific sense refers to the percentage of patients correctly classified by the test under evaluation. This percentage depends on the prevalence of the target disorder in the study group whenever sensitivity and specificity are not equal, and it weights false positive and false negative findings equally. Another single indicator is Youden's index [3,4]. It can be derived from sensitivity and specificity, and as such, it is independent of prevalence, but because it is a linear transformation of the mean sensitivity and specificity, its values are difficult to interpret [5].

The odds ratio used as single indicator of test performance is a third option. It is not prevalence dependent, and may be easier to understand, as it is a familiar epidemiologic measure.

The diagnostic odds ratio of a test is the ratio of the odds of positivity in disease relative to the odds of positivity in the nondiseased [6,7]. This means that the following relations hold:

$$\mathrm{DOR} = \frac{\mathrm{TP}}{\mathrm{FN}} \Big/ \frac{\mathrm{FP}}{\mathrm{TN}} = \frac{\mathrm{sens}}{(1-\mathrm{sens})} \Big/ \frac{(1-\mathrm{spec})}{\mathrm{spec}} \qquad (1)$$

Alternatively, the DOR can be read as the ratio of the odds of disease in test positives relative to the odds of disease in test negatives.

$$\mathrm{DOR} = \frac{\mathrm{TP}}{\mathrm{FP}} \Big/ \frac{\mathrm{FN}}{\mathrm{TN}} = \frac{\mathrm{PPV}}{(1-\mathrm{PPV})} \Big/ \frac{(1-\mathrm{NPV})}{\mathrm{NPV}} \qquad (2)$$

There also is a close relation between the DOR and likelihood ratios:

$$\mathrm{DOR} = \frac{\mathrm{TP}}{\mathrm{FP}} \Big/ \frac{\mathrm{FN}}{\mathrm{TN}} = \frac{\mathrm{LR}(+)}{\mathrm{LR}(-)} \qquad (3)$$

The value of a DOR ranges from 0 to infinity, with higher values indicating better discriminatory test performance. A value of 1 means that a test does not discriminate between patients with the disorder and those without it. Values lower than 1 point to improper test interpretation (more negative tests among the diseased). The inverse of the DOR can be interpreted as the ratio of negativity odds within the diseased relative to the odds of negativity within the nondiseased. The DOR rises steeply when sensitivity or specificity becomes near perfect, as is illustrated in Fig. 1 [6].

As can be concluded from above formulas, the DOR does not depend on the prevalence of the disease that the test is used for. Nevertheless, across clinical applications it is likely to depend on the spectrum of disease severity, as is the case for all other indicators of test performance [8,9]. Another point to consider is that, the DOR, as a global measure,

Table 2
Commonly used test indicators in diagnostic research

Test indicator	Formula	Definition
Sensitivity (true positive rate, TPR)	TP/(TP + FN)	Proportion positive test results among diseased
Specificity (true negative rate, TNR)	TN/(TN + FP)	Proportion negative test results among the "healthy"
Positive predictive value (PPV)	TP/(TP + FP)	Proportion diseased among subjects with a positive test result
Negative predictive value (NPV)	TN/(TN + FN)	Proportion nondiseased among subjects with a negative test result
Likelihood ratio of a positive test result (LR+)	sensitivity/(1−specificity)	Ratio of a positive test result among diseased to the same result in the "healthy"
Likelihood ratio of a negative test result (LR−)	(1−sensitivity)/specificity	Ratio of a negative test result among diseased to the same result in the "healthy"
Accuracy	(TP + TN)/(TP + TN + FP + FN)	Proportion correctly identified subjects
Youden's index	sensitivity + specificity−1	

Table 3
Comparison of three imaging tests in the diagnosis of peritoneal metastasis, of advanced ovarian cancer

	Sensitivity		Specificity		PPV		NPV		LR+		LR−		DOR	
Imaging test	%	(95% CI)	%	(95% CI)	%	(95% CI)	%	(95% CI)		(95% CI)		(95% CI)		(95% CI)
Ultrasound	69	(58–80)	93	(90–97)	78	(68–89)	90	(85–94)	10	(6.0–18)	0.33	(23–5.1)	31	(15–67)
CT scan	92	(84–100)	81	(75–87)	61	(50–72)	97	94–100	5.1	(3.6–6.9)	0.10	(0.038–3.2)	51	(17–151)
MRI	95	(89–100)	80	(73–87)	59	(47–71)	98	96–100	4.8	(3.4–6.7)	0.06	(0.016–3.7)	77	(18–340)

Reference standard: surgical and histopathologic findings [2]. Not all patients underwent all three imaging tests.
Prevalence of metastasis for ultrasound, CT scan, and MRI: respectively, 68/262, 50/212, and 41/175.

cannot be used to judge a test's error rates, at particular prevalences. Two tests with an identical DOR can have very different sensitivity and specificity, with distinct clinical consequences. If a 2 × 2 table contains zeroes, the DOR will be undefined. Adding 0.5 to all counts in the table is a commonly used method to calculate an approximation of the DOR [10,11]. Confidence intervals for range estimates and significance testing can be conventionally calculated with the following formula [12,13].

$$\mathrm{SE}(\log \mathrm{DOR}) = \sqrt{\frac{1}{\mathrm{TP}} + \frac{1}{\mathrm{TN}} + \frac{1}{\mathrm{FP}} + \frac{1}{\mathrm{FN}}} \tag{4}$$

A 95% confidence interval of the log DOR can then be obtained by:

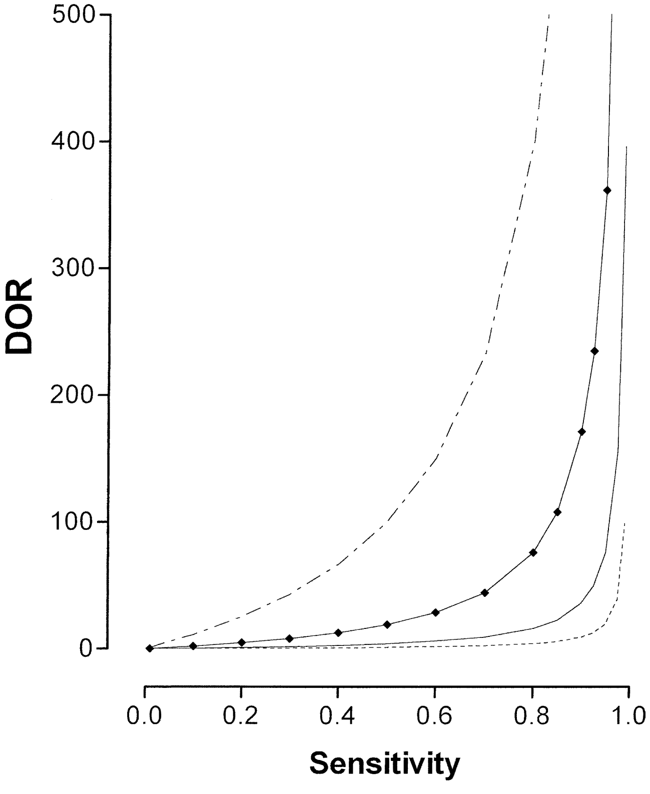

Fig. 1. Behavior of the odds ratio with changing sensitivity and specificity. Specificity: – · –· = 0.99, ● = 0.95, — = 0.80,- - - - = 0.50.

$$\log \mathrm{DOR} \pm 1.96\mathrm{SE}(\log \mathrm{DOR}) \tag{5}$$

Calculating the antilog of this expression (back-transformation) provides the confidence interval of the DOR.

In the example of the radiologic tests comparison, represented in Table 3, the estimated DOR for the ultrasound in detecting peritoneal metastases is 31 (0.69/(1 − 0.69))/((1 − 0.93)/0.93). This means that for the ultrasound the odds for positivity among subjects with peritoneal metastases is 31 times higher than the odds for positivity among subjects without peritoneal metastases. In the same way the DORs for the CT scan and magnetic resonance imaging (MRI) can be calculated (Table 3). MRI has the highest DOR in detecting peritoneal metastases compared to the ultrasound and CT scan (77 vs. 31 and 51, respectively). In contrast, ultrasound has the highest DOR in detecting liver metastases and lymph nodes: 54 vs. 17 for CT and 15 for MRI (Table 4).

If DORs had been presented in the original article, a quick comparison would have led to the conclusion that MR imaging performs best in diagnosing peritoneal metastases, whereas ultrasound does better in diagnosing lymph node and liver metastases in this population.

3. Polychotomous and continuous test outcomes

The performance of a test for which several cutoffs are available can be expressed by means of ROC analysis [14,15]. A receiver operating characteristic (ROC) curve plots the true positive rate on the Y-axis as a function of the false positive rate on the X-axis for all possible cutoff values of the test under evaluation. The area under the curve obtained (AUC) can subsequently be calculated as an alternative single indicator of test performance [16].

The AUC takes values between 0 and 1, with higher values indicating better test performance. These can be interpreted as an estimate of the probability that the test correctly ranks two individuals, of which one has the disease and one does not have the disease [16]. It can alternatively be interpreted as the average sensitivity across all possible specificities.

Table 4
Comparison of three imaging tests in the staging of ovarian cancer; lymph node, and hepatic metastases

	Lymph node metastases						Hepatic parenchymal metastase					
	Sensitivity		Specificity		DOR		Sensitivity		Specificity		DOR	
Imaging test	%	(95% Cl)	%	(95% Cl)	(95% Cl)		%	(95% Cl)	%	(95% Cl)	(95% Cl)	
Ultrasound	32	(11–52)	93	(89–96)	6.0	(2–18)	57	(20–94)	98	(96–99)	54	(9.9–299)
CT scan	43	(17–69)	89	(85–93)	6.1	(1.9–19)	40	(0–83)	96	(94–99)	17	(2.4–114)
MRI	38	(12–65)	84	(78–89)	3.2	(96–10)	40	(0–83)	96	(92–99)	15	(2.1–102)

Reference standard: surgical and histopathologic findings [2]. Lymph node metastasis: prevalence for ultrasound, CT scan, and MRI is, respectively, 19/255, 14/205, and 13/171. For hepatic parenchymal metastasis, 7/258, 5/212, and 5/165.

$$\mathrm{AUC} = \int_0^1 \frac{1}{1+\dfrac{1}{\mathrm{DOR}\cdot\left(\dfrac{x}{1-x}\right)}}\,dx \qquad (6)$$

If the DOR is constant for all possible cutoff values, the ROC curve will be symmetric (in relation to the diagonal $y = -x + 1$) and concave. In that case, a mathematical relation exists between the AUC and the DOR of a test (see formula 6). The higher the value of the DOR, the higher the AUC. AUCs of 0.85 and 0.90, for example, correspond to DORs of 13 and 24, respectively. An increase in AUC of 5% will almost double the DOR, which is a direct consequence of scale differences: the DOR has an upper limit of infinity, whereas the AUC takes values in the 0 to 1 range. For nonsymmetrical ROC curves the DOR is not constant over all cutoff points. In these cases the AUC cannot be calculated from the DOR associated with a single (or a few) cutoff values.

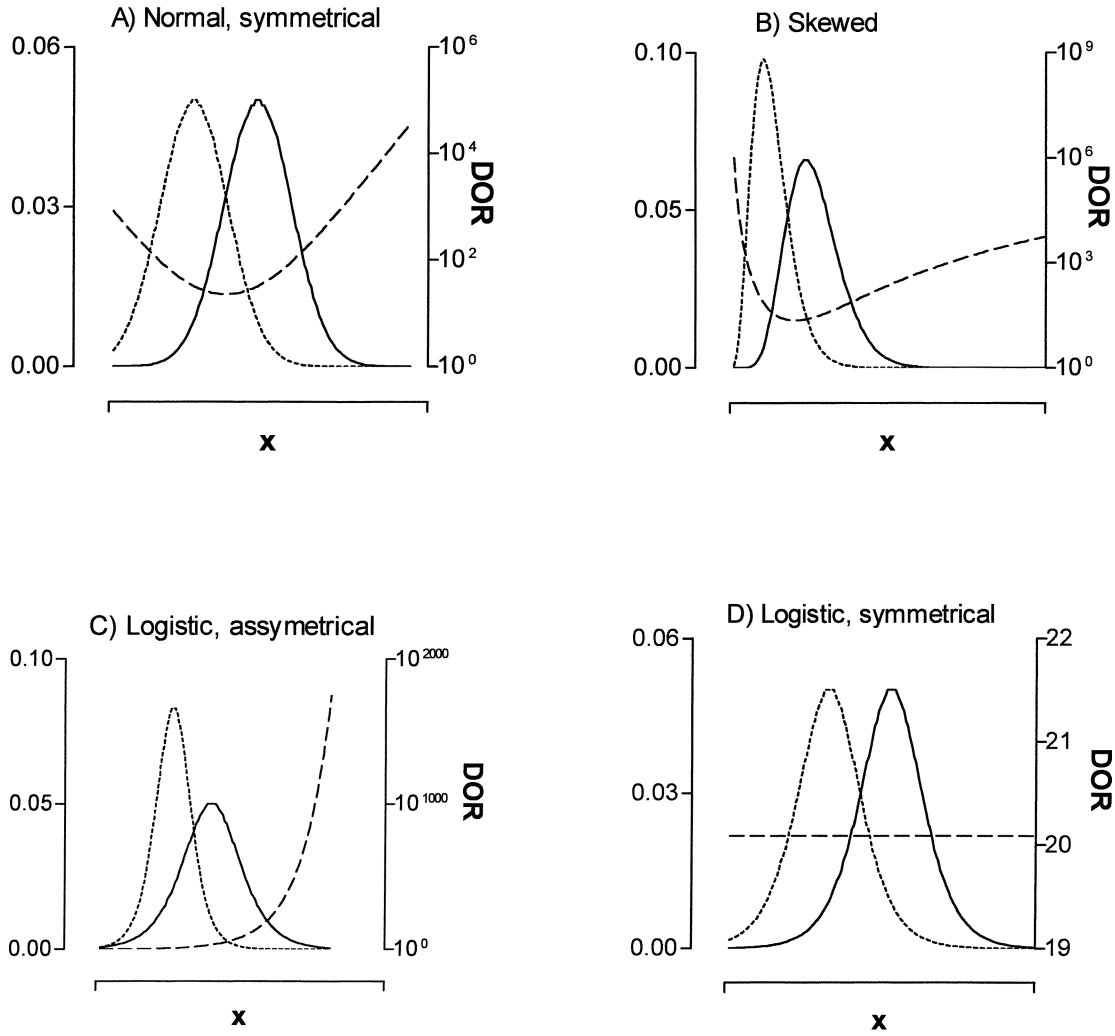

Fig. 2. Value of the DOR for all thresholds and distributions of test results in nondiseased and diseased. — — —: DOR: ·········: nondiseased population, ——: diseased population.

The shape of the ROC curve and the cutoff independence of the DOR depend on the underlying distribution of test results in patients with and without the target condition. Figure 2 shows several probability densities distributions of test results in diseased and nondiseased populations. It can be observed that the DOR is reasonably constant for a large range off cutoff points on the ROC curve, but for the extremes of sensitivity and specificity the DOR rises steeply. If the original or transformed results in both diseased and nondiseased follow a logistic distribution with equal SD, the DOR is constant for all possible cutoff values (Fig. 2D) [17].

4. The DOR in meta-analysis

The DOR offers considerable advantages in meta-analysis of diagnostic studies that combines results from different studies into summary estimates with increased precision. Meta-analysis of diagnostic tests offers statistical challenges, because of the bivariate nature of the conventional expressions of test performance. Simple pooling of sensitivity and specificity usually is inappropriate, as this approach ignores threshold differences [11,18]. In addition, heterogeneity may lead to an underestimation of a test's performance. The current strategy for meta-analysis, as endorsed by the Methods Working Group of the Cochrane Collaboration [19], builds on the methods described by Kardaun and Kardaun and Littenberg and Moses [11,20,21]. The approach by Littenberg and Moses relies on the linear regression of the logarithm of the DOR of a study (dependent variable) on a expression of the positivity threshold of that study (independent variable). If the regression line has a zero slope, the DOR is constant across studies. A summary ROC (sROC) can be produced after back-transforming the regression line. The resulting sROC will be symmetric and concave. In other words, study heterogeneity can be attributed to threshold differences. In the context of the DOR, the summary DOR of the study under evaluation can be obtained from the intercept ($e^{\text{intercept}}$) of the regression line [11,17]. Additional heterogeneity owing to variation in study characteristics (e.g., cohort vs. case–control) or clinical characteristics (e.g., heterogeneous prior therapy) can be evaluated simultaneously by adding these variables as covariates to the regression model, leaving a corrected estimated value for the pooled DOR. The resulting parameter estimates can be (back)transformed to relative diagnostic odds ratios (rDOR). An rDOR of 1 indicates that the particular covariate does not affect the overall DOR. A rDOR >1 means that studies, study centers, or patient subgroups with a particular characteristic have a higher DOR than studies without this characteristic. For a rDOR <1, the reverse holds [22]. When the DOR is homogeneous across studies, the DORs of different studies can also be pooled directly. Homogeneity can be tested by using the Q-test statistic or the H statistic [23].

We will illustrate the usefulness of the DOR in meta-analysis by reanalyzing a meta-analysis on the diagnostic performance of two magnetic resonance angiography techniques [3D gadolinium-enhanced (3D-GD) and 2D time of flight (2D-TOF)] detecting peripheral arteriosclerotic occlusive disease [24]. The separate meta-regression analysis yielded an intercept of 4.13 and a slope of 0.41 for 2D-TOF. For 3D-GD, these values were, respectively, 5.93 and −0.37. From the intercept the summary DORs can be calculated, respectively, $e^{4.13} = 62$ and $e^{5.93} = 376$. The nonzero slopes, indicated heterogeneity apart from threshold differences, which in turn, limits a direct comparison of summary odds ratios. To explore additional variation all studies of the two techniques were put together in one regression model. Then each available covariate was examined for its effect on the diagnostic performance. Most effect had the covariates 3D-GD vs. 2D-TOF technique, and a covariate dealing with the postprocessing technique (maximum intensity projections (MIP) in addition to transverse source images or multiplanar reformation (MIP+) vs. MIP alone). Subsequently, these two covariates were selected for the final multivariate model. The adjusted rDOR estimated was 7.5 (confidence interval: 2.8–22) for the 3D-GD and 4.5 (confidence interval: 1.5–14) for the use of MIP+. The confidence intervals did not contain the value 1. As such one can conclude that—after correction for heterogeneity—3D-GD and the use of MIP+ have a better diagnostic performance compared to 2D-TOF and the use of MIPs alone.

The methodology of systematic reviews and meta-analysis of diagnostic tests is still evolving, with new and potentially better methods being developed, better adapted to the inherently bivariate nature of the problem. Yet the convenience of the odds ratio in statistical modeling guarantees its future role in the meta-analysis of diagnostic tests.

5. Logistic regression

The DOR offers advantages when logistic regression is used with diagnostic problems. Logistic regression can be used to construct decision rules, reflecting the combined diagnostic value of a number of diagnostic variables [25]. Another application is the study of the added value of diagnostic tests [26]. With a single dichotomous test the logistic regression equation reads:

$$\mathrm{P(D|}x) = \frac{1}{1 + \exp^{-(\alpha + \beta x)}} \tag{7}$$

where x stands for the test result and the coefficients α and β have to be estimated. If a positive test result is coded as $x = 1$ and a negative as $x = 0$, we have

$$\mathrm{P(D|positive)} = \frac{1}{1 + \exp^{-(\alpha + \beta)}} \tag{8}$$

and

$$\mathrm{P(D|negative)} = \frac{1}{1 + \exp^{-\alpha}} \tag{9}$$

Next, one derives from expression 1, 8, and 9

$$\text{DOR} = [\text{P}(\text{D}|\text{positive})/(1 - \text{P}(\text{D}|\text{positive}))]/[\text{P}(\text{D}|\text{negative})/(1 - \text{P}(\text{D}|\text{negative}))] = \exp(\beta). \quad (10)$$

In other words, the DOR equals the regression coefficient, after exponentiation. Logistic regression modeling has been proposed as the preferred statistical method to obtain a posttest probability of disease when results from multiple tests are available. History taking and physical examination can also be considered as individual diagnostic tests. The posttest probability after having obtained test results x_1, x_2, … x_k is expressed as

$$\text{P}(\text{D}|X_1,X_2 \ldots X_k) = \frac{1}{1 + \exp^{-(\alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots + \beta_k x_k)}}. \quad (11)$$

With multiple dichotomous tests of which the results x_1, x_2 … x_k are coded as present (1) or absent (0), the corresponding coefficients β_1, β_2, … β_k equal the conditional logDOR [7]. These DORs are conditional: they depend on the other variables that have been used in the model [27]. If more information becomes available, a new regression equation has to be constructed to obtain the proper conditional DOR. This application is illustrated by a study that aimed to assess the value of symptoms in diagnosing arrhythmias in general practice [28]. The following equation from the logistic model was created to estimate the probability of arrhythmia in a patients with specific signs and symptoms: P(arrhythmia)=

$$\frac{1}{\left[1 + \exp^{-(-4.40 + 0.051*\text{age} + 0.47*\text{gender} + 1.11*\text{palpitations} + 0.78*\text{dyspnoea} + 0.45*\text{use of cardiovasmed.})}\right]}$$

Age is a continuous variable expressed in years. The odds ratio ($e^{0.051} = 1.05$), calculated from the respective coefficient, does not express the diagnostic performance of the variable age, but the OR for the increase in age per year. Gender is coded 1 for males and 0 for females. The use of cardiovascular medication, palpitations, and dyspnoea as recorded during consultation are coded as positive (1) or negative (0). Subsequently, the conditional DOR of each dichotomous variable, adjusted for the other variables, can be estimated. For gender, the DOR is $e^{0.47} = 1.6$, meaning that the odds for having arrhythmias is 1.6 times larger in males than in females. The adjusted odds ratios for palpitations during consultation, dyspnoea during consultation, and the use of cardiovascular medication are respectively 3.0, 2.2, and 1.6.

6. Discussion

The diagnostic odds ratio as a measure of test performance combines the strengths of sensitivity and specificity, as prevalence independent indicators, with the advantage of accuracy as a single indicator. These characteristics lend the DOR particularly useful for comparing tests whenever the balance between false negative and false positive rates is not of immediate importance. These features are also highly convenient in systematic reviews and meta-analyses.

In decisions on the introduction of a test in clinical practice, we are aware that the actual balance between the true positive rate and false positive rate often matters [29]. Whenever false positives and false negatives are weighted differentially, both the prevalence and the conditional error rates of the test have to be taken into consideration to make a balanced decision. In these cases, the DOR is less useful, as it does not distinguish between the two types of diagnostic mistake. If ruling-out or ruling-in of the target condition is the primary intended use of a test, conditional indicators such as sensitivity and specificity still have to be used.

As all available measures of test performance, the DOR of a test is unlikely to be a test-specific constant. Its magnitude likely depends on the spectrum of disease as well as on preselection through the use of other tests [6,30]. Despite this universal caveat for indicators of diagnostic tests, we feel that a more systematic use of the odds ratio in diagnostic research can contribute to more consistent applications of diagnostic knowledge.

Some may object that there are already too many indicators of test performance. With such an abundance of choices, there is little need for yet another statistic. This may be true, but it is hard to see how the selection can or should be produced. Each of the indicators serves a different purpose. Sensitivity and specificity are expressions of the conditional hit rates of the test. Predictive values or posterior probabilities are the numbers that are most salient for clinical practice. The so-called likelihood ratios come in handy for comparing the diagnostic content of multiple possible test results and for transforming those into posttest probabilities. Among those helpful indicators, the diagnostic odds ratio has a place as a single statistic with a long history and useful statistical properties.

Acknowledgments

The authors thank Dr. Guus Hart for critically reviewing and giving useful suggestions on earlier drafts.

References

[1] Sackett DL, Haynes RB, Guyatt GH, Tugwell P. Clinical epidemiology: a basic science for clinical medicine. Boston: Little, Brown and Co.; 1991.

[2] Tempany CM, Zou KH, Silverman SG, Brown DL, Kurtz AB, McNeil BJ. Staging of advanced ovarian cancer: comparison of imaging modalities-report from the Radiological Diagnostic Oncology Group. Radiology 2000;215(3):761–7.

[3] Youden WJ. Index for rating diagnostic tests. Cancer 1950;3:32–5.

[4] Hilden J, Glasziou P. Regret graphs, diagnostic uncertainty and Youden's Index. Stat Med 1996;15(10):969–86.

[5] Linnet K. A review on the methodology for assessing diagnostic tests. Clin Chem 1988;34(7):1379–86.

[6] Kraemer HC. Risk ratios, odds ratio, and the test QROC. In: Evaluating medical tests. Newbury Park, CA: SAGE Publications, Inc.; 1992. p. 103–13.

[7] Korte de PJ. Probability analysis in diagnosing coronary artery disease. Thesis, Universitaire Pers Maastricht, 1993.

[8] Hlatky MA, Pryor DB, Harrell FE Jr, Califf RM, Mark DB, Rosati RA. Factors affecting sensitivity and specificity of exercise electrocardiography. Multivariable analysis. Am J Med 1984;77(1):64–71.

[9] Moons KG, van Es GA, Deckers JW, Habbema JD, Grobbee DE. Limitations of sensitivity, specificity, likelihood ratio, and Bayes' theorem in assessing diagnostic probabilities: a clinical example. Epidemiology 1997;8(1):12–7.

[10] Haldane J. The estimation and significance of the logarithm of a ratio of frequencies. Ann Hum Genet 1955;20:309–14.

[11] Littenberg B, Moses LE. Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method. Med Decis Making 1993;13(4):313–21.

[12] Altman D. Comparing groups-categorical data. In: Altman D, editor. Practical statistics for medical research. Florida: Chapman&Hall/CRC; 1997. p. 229–76.

[13] Bland JM, Altman DG. The odds ratio. BMJ 2000;320:1468.

[14] Choi BCK. Slopes of a receiver operating characteristic curve and likelihood ratios for a diagnostic lest. Am J Epidemiol 1998;148(11): 1127–32.

[15] Swets JA. The relative operating characteristic in psychology. Science 1973;182:990–1000.

[16] Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143(1): 29–36.

[17] Hasselblad V, Hedges LV. Meta-analysis of screening and diagnostic tests. Psychol Bull 1995;117(1):167–78.

[18] Irwig L, Macaskill P, Glasziou P, Fahey M. Meta-analytic methods for diagnostic test accuracy. J Clin Epidemiol 1995;48(1)(1):119–30.

[19] The Methods Working Group. On systematic review of screening and diagnostic tests: recommended methods. 1995; http://www.cochrane.org.au/.

[20] Kardaun JW, Kardaun OJ. Comparative diagnostic performance of three radiological procedures for the detection of lumbar disk herniation. Methods Inf Med 1990;29(1)(1):12–22.

[21] 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(14):1293–316.

[22] Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, Bossuyt PM. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999;282(11):1061–6.

[23] Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med 2002;21(11):1539–58.

[24] Nelemans PJ, Leiner T, de Vet HC, van Engelshoven JM. Peripheral arterial disease: meta-analysis of the diagnostic performance of MR angiography. Radiology 2000;217(1)(1):105–14.

[25] Wells PS, Anderson DR, Rodger M, Ginsberg JS, Kearon C, Gent M, Turpie AG, Bormanis J, Weitz J, Chamberlain M, Bowie D, Barnes D, Hirsh J. Derivation of a simple clinical model to categorize patients probability of pulmonary embolism: increasing the models utility with the SimpliRED D-dimer. Thromb Haemost 2000;83(3): 416–20.

[26] Clark TJ, Bakour SH, Gupta JK, Khan KS. Evaluation of outpatient hysteroscopy and ultrasonography in the diagnosis of endometrial disease. Obstet Gynecol 2002;99(6):1001–7.

[27] Knottnerus JA. Application of logistic regression to the analysis of diagnostic data: exact modeling of a probability tree of multiple binary variables. Med Decis Making 1992;12(2):93–108.

[28] Zwietering PJ, Knottnerus JA, Rinkens PE, Kleijne MA, Gorgels AP. Arrhythmias in general practice: diagnostic value of patient characteristics, medical history and symptoms. Fam Pract 1998;15(4): 343–53.

[29] Glasziou P, Hilden J. Test selection measures. Med Decis Making 1989;9(2):133–41.

[30] Hlatky MA, Lee KL, Botvinick EH, Brundage BH. Diagnostic test use in different practice settings. A controlled comparison. Arch Intern Med 1983;143(10):1886–9.