Journal of Biometrics & Biostatistics

Okeh and Okeh, J Biomet Biostat 2012, 3:1
DOI: 10.4172/2155-6180.1000132

Review Article Open Access

Evaluating Measures of Indicators of Diagnostic Test Performance: Fundamental Meanings and Formulars

Okeh UM* and Okoro CN

Department of Industrial Mathematics and Applied Statistics, Ebonyi State University Abakaliki, Nigeria

Abstract

Diagnostic accuracy relates to the ability of a test to discriminate between the target condition and health. This discriminative potential can be quantified by the measures of diagnostic accuracy such as sensitivity and specificity, predictive values, likelihood ratios, error rates, the area under the ROC curve, Youden's index and diagnostic odds ratio. Different measures of diagnostic accuracy relate to the different aspects of diagnostic procedure: while some measures are used to assess the discriminative property of the test, others are used to assess its predictive ability. Measures of diagnostic accuracy are not fixed indicators of a test performance, some are very sensitive to the disease prevalence, while others to the spectrum and definition of the disease. Furthermore, measures of diagnostic accuracy are extremely sensitive to the design of the study. Studies not meeting strict methodological standards usually over- or under-estimate the indicators of test performance as well as they limit the applicability of the results of the study. STARD initiative was a very important step toward the improvement of the quality of reporting of studies of diagnostic accuracy. STARD statement should be included into the Instructions to authors by scientific journals and authors should be encouraged to use the checklist whenever reporting their studies on diagnostic accuracy. Such efforts could make a substantial difference in the quality of reporting of studies of diagnostic accuracy and serve to provide the best possible evidence to the best for the patient care. This brief review outlines some basic definitions, formulas and characteristics of the measures of diagnostic accuracy.

Keywords: Diagnostic accuracy; Sensitivity; Specificity; Likelihood ratio; DOR; AUC; Predictive values

Introduction

There is no single statistic that can adequately represent the agreement between a diagnostic test and a reference standard. Many different statistics have a part to play in the analysis of such studies. This discriminative ability can be quantified by the measures of diagnostic accuracy: Sensitivity, Specificity, Receiver operating characteristic curve (ROC curve), Likelihood ratio (LR) for positive test, Likelihood ratio (LR) for negative test, Odds ratio (OR), Matthew's correlation coefficients(MCC), False discovery rate(FDR), Positive predictive value (PPV), Youden's index, Negative predictive value (NPV),Error rates and Confidence interval. Diagnostic accuracy of any diagnostic procedure or a test gives us an answer to the following question: "How well this test discriminates between certain two conditions of interest (health and disease; two stages of a disease etc.)?". Different measures of diagnostic accuracy relate to the different aspects of diagnostic procedure. Some measures are used to assess the discriminative property of the test, others are used to assess its predictive ability [1]. While discriminative measures are mostly used by health policy decisions, predictive measures are most useful in predicting the probability of a disease in an individual [2]. Furthermore, it should be noted that measures of a test performance are not fixed indicators of a test quality and performance. Measures of diagnostic accuracy are very sensitive to the characteristics of the population in which the test accuracy is evaluated. Some measures largely depend on the disease prevalence, while others are highly sensitive to the spectrum of the disease in the studied population. It is therefore of utmost importance to know how to interpret them as well as when and under what conditions to use them.

Diagnostic Effectiveness

Another global measure of diagnostic accuracy is so called diagnostic accuracy (effectiveness), expressed as a proportion of correctly classified subjects (TP+TN) among all subjects (TP+TN+FP+FN). Diagnostic accuracy is affected by the disease prevalence. With the same sensitivity and specificity, diagnostic accuracy of a particular test increases as the disease prevalence decreases. This data, however, should be handled with care. In fact, this does not mean that the test is better if we apply it in a population with low disease prevalence. It only means that in absolute number the test gives more correctly classified subjects. This percentage of correctly classified subjects should always be weighed considering other measures of diagnostic accuracy, especially predictive values. Only then a complete assessment of the test contribution and validity could be made.

Consider table 1 below which indicates the test status and disease condition of patient from where some common indicators of test performance will be derived. Some of these indicators are the sensitivity of the test, its specificity, the positive and negative predictive values, Odds ratio, error rates and the positive and negative likelihood ratios [3].

A frequent application of Bayes' theorem is in evaluating the performance of a diagnostic test intended for use in a screening program. From table 1 above, let B denote the event that a person has the disease in question; $\overline{B}$ the event that he does not have the disease; let A be the event that he gives a positive response to the test; and $\overline{A}$ the event that he gives a negative response. Let P be the Prevalence of the disease. Prevalence is the number of cases of a disease that are present in

***Corresponding author:** Okeh UM, Department of Industrial Mathematics and Applied Statistics, Ebonyi State University Abakaliki, Nigeria, E-mail: umokeh1@yahoo.com

Received December 01, 2011; **Accepted** January 12, 2012; **Published** January 16, 2012

Citation: Okeh UM, Okoro CN (2012) Evaluating Measures of Indicators of Diagnostic Test Performance: Fundamental Meanings and Formulars. J Biomet Biostat 3:132. doi:10.4172/2155-6180.1000132

a particular population at a given time [4]. Based on the above table, it is given by P=mean (p_i) while Q is the level of the test given by Q=mean (q_i). But P'=1-P and Q' =1-Q. From this Table 1, TP is true positive, FN is false positive, TN is true negative and FP is false positive while N is the total number of patients/subject considered. The first step in the calculation of sensitivity and specificity is to make a 2x2 table with groups of subjects divided according to a gold standard or (reference method) in columns, and categories according to test in rows as seen in Table 1 above.

Sensitivity and Specificity

The first step in the calculation of sensitivity and specificity is to make a 2x2 table with groups of subjects divided according to a gold standard or (reference method) in columns, and categories according to test in rows (Table 1).

Let π and θ be sensitivity and specificity of the tests respectively. The results of this trial of the screening test may be represented by the two conditional probabilities $P(A|B)$ and $P(\bar{A}|\bar{B})$. Sensitivity is expressed in percentage and defines the proportion of true positive subjects with the disease in a total group of subjects with the disease (TP/TP+FN). Actually, sensitivity is defined as the probability of getting a positive test result in subjects with the disease (T+|B+). Hence, it relates to the potential of a test to recognize subjects with the disease. The sensitivity (π) given as $P(A|B)$ is the conditional probability of a positive response given that the person has the disease; the larger is $P(A|B)$, the more sensitive is the test. From table 1, conditional probability in terms of sensitivity is also given by

$$P(T=1|D=1)=\Pr(T+|D+)=\frac{TP}{TP+FN} \tag{1}$$

While $P(\bar{A}|\bar{B})$ is the conditional probability of a negative response given that the person is free of the disease; the smaller is $P(\bar{A}|\bar{B})$, the more specific is the test. We can also define that 1-π = $P(T=0|D=1)$. Specificity is a measure of a diagnostic test accuracy, complementary to sensitivity. It is defined as a proportion of subjects without the disease with negative test result in total of subjects without disease (TN/TN+FP). In other words, specificity represents the probability of a negative test result in a subject without the disease (T-|B-). Therefore, we can postulate that specificity relates to the aspect of diagnostic accuracy that describes the test ability to recognize subjects without the disease, i.e. to exclude the condition of interest. From table 1 also, specificity is given by

$$P(T=0|D=0)=\Pr(T-|D-)=\frac{TN}{FP+TN} \tag{2}$$

Also 1-θ = $P(T=1|D=0)$. Neither sensitivity nor specificity is not influenced by the disease prevalence, meaning that results from one study could easily be transferred to some other setting with a different prevalence of the disease in the population. Nonetheless, sensitivity and specificity can vary greatly depending on the spectrum of the disease in the studied group. Worthy of note also is the predictive values which includes the positive predictive values (PPV) and negative predictive values (NPV). Positive predictive value (PPV) defines the probability of having the state/disease of interest in a subject with positive result (B+|T+). Therefore PPV represents a proportion of patients with positive test result in total of subjects with positive result (TP/TP+FP). PPV is defined as the probability of a positive diagnosis when the test is positive. Therefore, PPV represents a proportion of patients with positive test result in total of subjects with positive result. It is also seen as the ratio of the number of true positives to the total number of positive tests [5]. In terms of Bayes formula, it is given by

$$PPV=P(D+|T+)=\frac{P(T+|D+)P(D+)}{P(T+)}=(SE.P)/Q=TP/Q \tag{3}$$

where Q=TP+FP=level of the test. Negative predictive value (NPV) describes the probability of not having a disease in a subject with a negative test result (B-|T-). NPV is defined as a proportion of subjects without the disease with a negative test result in total of subjects with negative test results (TN/TN+FN). Also the NPV in terms of Bayes formula is the probability of negative diagnosis when the test is negative or the proportion of subjects without the disease with a negative test result in total of subjects with negative test results and it is given by

$$NPV=P(D-|T-)=\frac{P(T-|D-)P(D-)}{P(T-)}$$
$$=(SP.P')/Q'=TN/Q' \text{ where } Q'=TN+FN \tag{4}$$

According to the table 1 above for conditional probability, the predictive values are defined mathematically as:

$$PPV=\frac{TP}{TP+FP}=P(D+|T+) \text{ or } P(T+|D+) \tag{5}$$

$$\text{Also } NPV=\frac{TN}{FN+TN}=P(D-|T-) \text{ or } P(T-|D-) \tag{6}$$

Also, the two can be defined as PPV = Pr(D=1|T=1) and NPV = Pr(D=0|T=0). Unlike sensitivity and specificity, predictive values are largely dependent on disease prevalence in examined population. Unlike sensitivity and specificity, predictive values are largely dependent on disease prevalence in examined population. Therefore, predictive values from one study should not be transferred to some other setting with a different prevalence of the disease in the population. Prevalence affects PPV and NPV differently. PPV is increasing, while NPV decreases with the increase of the prevalence of the disease in a population. Whereas the change in PPV is more substantial, NPV is somewhat weaker influenced by the disease prevalence. Therefore, predictive values from one study should not be transferred to some other setting with a different prevalence of the disease in the population. Prevalence affects PPV and NPV differently. PPV is increasing, while NPV decreases with the increase of the prevalence of the disease in a population. Whereas the change in PPV is more substantial, NPV is somewhat weaker influenced by the disease prevalence. It is important to say that there exist error rates to be expected if the test is actually used in a screening program. If a positive result is taken to indicate the presence of the disease, then the false positive rate (FPR) or say P_{F+} is the proportion of people, among those responding positive, who are actually free of the disease. It can also be denoted as $P(\bar{B}|A)$. According to Bayes' theorem and based on the Table 1,

$$P_{F+}=P(\bar{B}|A)=P(T=1|D=0)=1-\theta=$$
$$\frac{P(A|\bar{B})P(\bar{B})}{P(A)}=\frac{P(A|\bar{B})(1-P(B))}{P(A)} \tag{7}$$

$$\text{Since } P(\bar{B})=1-P(B).$$

The false negative rate (FNR) or say P_{F-}, is the proportion of people, among those responding negative on the test, who nevertheless have the disease, or $P(\bar{B}|A)$. Again by Bayes' theorem,

Hypothesis	Accept	Reject	Total
Null true	U	V	m_0
Alternative true	T	S	m_1
	W	R	m

Table 1: Outcomes when testing m hypothesis.

$$P_{F-} = P(B|\overline{A}) = 1-\pi = P(T=0|D=1) = \frac{P(\overline{A}|B)P(B)}{P(\overline{A})} = \frac{(1-P(A|B))P(B)}{1-P(A)} \qquad (8)$$

$Since\ P(\overline{A}|B) = 1-P(A|B)\ and\ P(\overline{A}) = 1-P(A).$

Looking at the mathematical definitions of FPR and FNR above, one discovers that P (A) and P(B) has to be clearly defined so as to make the formulas complete.

Derivation of the Error Rates Formulas

A perfect diagnostic procedure has the potential to completely discriminate subjects with and without disease. Values of a perfect test which are above the cut-off are always indicating the disease, while the values below the cut-off are always excluding the disease. Unfortunately, such perfect test does not exist in real life and therefore diagnostic procedures can make only partial distinction between subjects with and without disease. Values above the cut-off are not always indicative of a disease since subjects without disease can also sometimes have elevated values. Such elevated values of certain parameter of interest are called false positive values (FP). On the other hand, values below the cut-off are mainly found in subjects without disease. However, some subjects with the disease can have them too. Those values are false negative values (FN). Therefore, the cut-off divides the population of examined subjects with and without disease in four subgroups considering parameter values of interest. According to table 1 above, we define as follows the following terms:

1. True positive (TP) –subjects with the disease with the value of a parameter of interest above the cut-off
2. False positive (FP) –subjects without the disease with the value of a parameter of interest above the cut-off
3. True negative (TN) –subjects without the disease with the value of a parameter of interest below the cut-off
4. False negative (FN) –subjects with the disease with the value of a parameter of interest below the cut-off

Let us consider $P(A)$ and $P(B)$ for this evaluation based on table 1 above. If

$$P(A) = \frac{N_A}{N} = \frac{N_{AB}+N_{A\overline{B}}}{N} = \frac{N_{AB}}{N} + \frac{N_{A\overline{B}}}{N} \qquad (9)$$

and N_A indicates the total number of people who test positive, then N_{AB} denotes the number of people who have the disease and respond positive while N_{AB} denotes the number of people who are free of the disease and respond positive. Multiplying and dividing the first of the two terms on the right-hand side of the above equation by N_B, the number of people with the disease, we find that

$$\frac{N_{AB}}{N} = \frac{N_{AB}}{N_B}\frac{N_B}{N} = P(A|B)P(B). \qquad (10)$$

Similarly, by multiplying and dividing the second term by N_B, the number of people without the disease, we find that

$$\frac{N_{A\overline{B}}}{N} = \frac{N_{A\overline{B}}}{N_{\overline{B}}}\frac{N_{\overline{B}}}{N} = P(A|\overline{B})P(\overline{B}). \qquad (11)$$

Substituting the expressions from the above last three equations in $P(A)$ defined above, we find that

$$P(A) = P(A|B)P(B) + P(A|\overline{B})P(\overline{B}). \qquad (12)$$

This equation is a special case of the familiar result that an overall rate-$P(A)$ is a weighted average of specific rates- $P(A|B)\ and\ P(A|\overline{B})$ -with the weights being the proportions of people in the specific categories $P(B)\ and\ P(\overline{B})$. Since $P(\overline{B}) = 1-P(B)$, then the above equation becomes

$$\begin{aligned} P(A) &= P(A|B)P(B) + P(A|\overline{B})(1-P(B)) \\ &= P(A|\overline{B}) + P(B)(P(A|B) - P(A|\overline{B})). \end{aligned} \qquad (13)$$

Substituting this equation in the equation of P_{F+} above yields an expression for the FPR thus,

$$P_{F+} = \frac{P(A|\overline{B})(1-P(B))}{P(A|\overline{B}) + P(B)(P(A|B) - P(A|\overline{B}))}. \qquad (14)$$

Also substituting for $P(A)$ in the equation of P_{F-} above yields the expression for the FNR or

$$P_{F-} = \frac{(1-P(A|B))P(B)}{1-P(A|\overline{B}) - P(B)(P(A|B) - P(A|\overline{B}))}. \qquad (15)$$

In conclusion, the two error rates are functions of the proportions $P(A|B)\ and\ P(A|\overline{B})$ which may be estimated from the results of a trial of the screening test; and of the overall case rate $P(B)$, for which an accurate estimate is rarely available.

Likelihood Ratios

Likelihood ratio is a very useful measure of diagnostic accuracy. It is defined as the ratio of expected test result in subjects with a certain state/disease to the subjects without the disease. From a clinical standpoint, a diagnostic test should give a sense of how more or less likely the disease being tested for is present or not, i.e., does the result of the diagnostic test change probability of the disease being present or not?. Likelihood ratios can quantify the change in the probability of disease given the results of a diagnostic test. Likelihood ratios are alternative statistics for summarizing diagnostic accuracy, which have several particularly powerful properties that make them more useful clinically than other statistics [6]. Each test result has its own likelihood ratio, which summarizes how many times more (or less) likely patients with the disease are to have that particular result than patients without the disease. More formally, it is the ratio of the probability of the specific test result in people who do have the disease to the probability in people who do not. A likelihood ratio greater than 1 indicates that the test result is associated with the presence of the disease, whereas a likelihood ratio less than 1 indicates that the test result is associated with the absence of disease. The further likelihood ratios are from 1 the stronger the evidence for the presence or absence of disease. Likelihood ratios above 10 and below 0.1 are considered to provide strong evidence to rule in or rule out diagnoses respectively in most circumstances [7]. In other words, it indicates large, often clinically significant differences. A likelihood ratio of 1 implies that there will be no difference between pretest and posttest probabilities. In other words, the two ratios are equal, such that the test is of no value. Likelihood ratios between 1 and 2 and between 0.5 and 1 indicate small differences (rarely clinically significant). When tests report results as being either positive or negative the two likelihood ratios are called the positive likelihood ratio and the negative likelihood ratio. It is vital to note that sensitivity and specificity are combined into one to have the likelihood ratio (LR). The likelihood ratio is defined as: “The probability of a subject with the disease having the test result divided by the probability of the subject without the disease having that same test result”. From table 1 above

$$LR = \frac{\Pr obability\ of\ result\ in\ diseased\ persons}{\Pr obability\ of\ result\ in\ nondiseased\ persons} \qquad (16)$$

When test results are dichotomized, every test has two likelihood ratios, one corresponding to a positive test and that of negative test.

Positive test likelihood ratio (LR$^+$) tells us how much more likely the positive test result is to occur in subjects with the disease compared to those without the disease. It is usually higher than 1 because it is more likely that the positive test result will occur in subjects with the disease than in subject without the disease. LR+ is the best indicator for ruling-in diagnosis. The higher the LR+ the test is more indicative of a disease. Good diagnostic tests have LR+ > 10 and their positive result has a significant contribution to the diagnosis. LR+ can be simply calculated according to the following formulas:

$$LP+=\frac{\Pr obability\ that\ test\ is\ positive\ in\ diseased\ persons}{\Pr obability\ that\ test\ is\ positive\ in\ non-diseased\ persons}$$

$$LP+=\frac{Sensitivity}{1-Specificity}=\frac{\Pr(T+|D+)}{\Pr(T+|D-)}=\frac{TP/TP+FN}{FP/FP+TN} \qquad (17)$$

Likelihood ratio for negative test result (LR-) represents the ratio of the probability that a negative result will occur in subjects with the disease to the probability that the same result will occur in subjects without the disease. Therefore, LR- tells us how much less likely the negative test result is to occur in a patient than in a subject without disease. LR- is usually less than 1 because it is less likely that negative test result occurs in subjects with than in subjects without disease. LR- is calculated according to the following formulas:

$$LP-=\frac{\Pr obability\ that\ test\ is\ negative\ in\ diseased\ persons}{\Pr obability\ that\ test\ is\ negative\ in\ non-diseased\ persons}$$

$$=\frac{1-Sensitivity}{Specificity}=\frac{\Pr(T-|D+)}{\Pr(T-|D-)}=\frac{FN/TP+FN}{TN/FP+TN} \qquad (18)$$

LR- is a good indicator for ruling-out the diagnosis. Good diagnostic tests have LR- < 0,1. The lower the LR- the more significant contribution of the test is in ruling-out, i.e. in lowering the posterior probability of the subject having the disease. Since both specificity and sensitivity are used to calculate the likelihood ratio, it is clear that neither LR+ nor LR- depend on the disease prevalence in examined groups. Consequently, the likelihood ratios from one study are applicable to some other clinical setting, as long as the definition of the disease is not changed. If the way of defining the disease varies, none of the calculated measures will apply in some other clinical context. Meanwhile, it is also important to define mathematically the following terms as it relates to table 1 above.

$$\Pr etest\ probability(\Pr evalence)=\frac{\Pr etest\ probability}{1-\Pr etest\ probability}=\frac{TP+FN}{TP+FP+TN+FN} \qquad (19)$$

$$\Pr etest\ Odds=\frac{\Pr evalence}{1-\Pr evalence}=\frac{TP+FN}{FP+TN} \qquad (20)$$

$$Posttest\ probability=\frac{Post\ test\ Odds}{1+\ post\ test\ odds} \qquad (21)$$

$$Posttest\ Odds=\Pr etest\ Odds\times Likelihood\ ratio \qquad (22)$$

LR directly links the pre-test and post-test probability of a disease in a specific patient [8]. Simplified, LR tells us how many times more likely particular test result is in subjects with the disease than in those without disease. Likelihood ratios provide an estimation of whether there will be significant change in pretest to posttest probability of a disease given the test result, and thus can be used to make quick estimates of the usefulness of contemplated diagnostic tests in particular situations. The simplest method for calculating posttest probability from pretest probability and likelihood ratios is to use a nomogram. The clinician places a straightedge through the points that represent the pretest probability and the likelihood ratio and then reads the posttest probability where the straightedge crosses the posttest probability line. A more formal way of calculating posttest probabilities uses the likelihood ratio as follows: Pretest odds × Likelihood ratio = Posttest odds.

Odds Ratio

The Odds ratio or diagnostic odds ratio is the probability of the presence of a disease of a specific disease divided by the probability of its absence. Diagnostic odds ratio is also one global measure for diagnostic accuracy, used for general estimation of discriminative power of diagnostic procedures and also for the comparison of diagnostic accuracies between two or more diagnostic tests. DOR of a test is the ratio of the odds of positivity in subjects with disease relative to the odds in subjects without disease [9]. It is calculated according to the formula: DOR = (TP/FN)/(FP/TN). DOR depends significantly on the sensitivity and specificity of a test. A test with high specificity and sensitivity with low rate of false positives and false negatives has high DOR. With the same sensitivity of the test, DOR increases with the increase of the test specificity. For example, a test with sensitivity > 90% and specificity of 99% has a DOR greater than 500. The Odds ratio reflects the prevalence of the disease in a population. For example, the odds are 1:4 for finding a disease in a population with a 20% probability of occurrence (prevalence). Also odds of 3:1 in favor of the first outcome means that the first outcome occurs 3 times for each single occurrence of the second outcome. Similarly, odds of 5:2 means that the first outcome occurs 5 times for each 2 occurrences of the second outcome. To use this formulation above, probabilities must be converted to odds, where the odds of having a disease are expressed as the chance of having the disease divided by the chance of not having the disease.

$$\text{Recall that } Odd=\frac{\Pr obability}{1-\Pr obability}, \qquad (23)$$

$$\Pr obability=\frac{Odds}{1+odds}\ \text{and Probability}=\frac{\mathrm{A}}{\mathrm{A+B}} \qquad (24)$$

when odds are expressed as a:b. To estimate the potential benefit of a diagnostic test, the clinician first estimates the pretest odds of disease given all available clinical information and then multiplies the pretest odds by the positive and negative likelihood ratios. The results are the posttest odds, or the odds that the patient has the disease if the test is positive or negative. To obtain the posttest probability, the odds are converted to a probability as seen above. Meanwhile, odds ratio is a measure of effect size, describing the strength of association or non independence between two binary data values [10,11,12]. It is used as a descriptive statistic, and plays an important role in logistic regression. Unlike other measures of association for paired binary data such as the relative risk, the odds ratio treats the two variables being compared symmetrically, and can be estimated using some types of non-random samples.

Derivation of Odds Ratio Formular for AD/BC

From table 1 above where we have the test result status and disease status categorized/classified into A,B,C and D. We shall have the following:

$$Odds\ of\ exp\ osure\ in\ those\ with\ the\ disease=$$

$$\frac{\Pr obability\ of\ \exp osure\ in\ those\ with\ disease}{1-\Pr obability\ of\ \exp osure\ \mathrm{in}\ \mathrm{t}hose\ with\ disease} \qquad (25)$$

$$\text{But, }\Pr\ ob\ of\ exposure\ in\ those\ with\ the\ disease=\frac{A}{A+C}$$

Therefore, *the Odds of exposure in those with the disease =*

$$\frac{A}{A+C}/1-\frac{A}{A+C}=\frac{A}{A+C}/\frac{C}{A+C}=\frac{A}{C} \qquad (26)$$

Similarly, *Odds of exposure in those without the disease* =

$$\frac{\Pr obability\ of\ \exp osure\ in\ those\ without\ disease}{1-\Pr obability\ of\ \exp osure\ in\ t\,hose\ without\ disease} \qquad (27)$$

But, Pr *obability of exposure in those without disease* = $\dfrac{B}{B+D}$

So odds of exposure in those without the disease =

$$\frac{B}{B+D}/1-\frac{B}{B+D}=\frac{B}{B+D}/\frac{D}{B+D}=\frac{B}{D} \qquad (28)$$

$Since, 1-\dfrac{B}{B+D}=\dfrac{B+D-B}{B+D}=\dfrac{D}{B+D}$, Based on the above contingency table, the odds ratio is interpreted as

$$Odds\ ratio=\frac{Odds\ of\ test\ result\ with\ disease}{Odds\ of\ same\ result\ without\ disease} \qquad (29)$$

Or $Odds\ ratio=\dfrac{Odds\ of\ \exp osure\ \ in\ t\,hose\ with\ the\ disease}{Odds\ of\ \exp osure\ in\ those\ without\ the\ disease}$

$$Odds(D+)=\frac{\Pr obability(D+)}{1-\Pr obability(D+)} \qquad (30)$$

where $\Pr(D+)=\dfrac{Odds(D+)}{1+Odds(D+)}$ (31)

$$\text{In general, Odds ratio}=\frac{(A/C)}{(B/D)}=AD/BC \qquad (32)$$

Definition of Odds Ratio in Terms of Group-Wise Odds

The odds ratio is the ratio of the odds of an event occurring in one group to the odds of it occurring in another group. The term is also used to refer to sample-based estimates of this ratio. This group in case of this study is a dichotomous classification. If the probabilities of the event in each of the groups are p_1 (first group) and p_2 (second group), then the odds ratio is: $\dfrac{p_1/(1-p_1)}{p_2/(1-p_2)}=\dfrac{p_1/q_1}{p_2/q_2}=\dfrac{p_1q_2}{p_2q_1}=\dfrac{n_{11}/n_{12}}{n_{21}/n_{22}}=\dfrac{n_{11}n_{22}}{n_{21}n_{12}}$ (33),

where $q_x = 1 - p_x$. This is the odds ratio for comparing two proportions. According to Glas et al [9], OR is based on likelihood and it is given by

$$\text{Diagnostic odds ratio}=\frac{likelihood\ ratio+}{likelihood\ ratio\ ratio-} \qquad (34)$$

It is vital to note the following points about OR. OR ranges from 0 to infinity with higher values indicating better discriminating test performance. A value of 1 means that a test does not discriminate between patients with the disorder and those without it. That is it shows no difference in risk of group 1 compared to 2. In other words, an odds ratio of 1 indicates that the condition or event under study is equally likely to occur in both groups. The DOR does not depend on the prevalence of the disease [9]. It tends to be skewed (not symmetric). If OR >1, it indicates an increased risk of group 1 compared to 2. This means that an odds ratio greater than 1 indicates that the condition or event is more likely to occur in the first group. If OR is less than 1 it indicates that the condition or event is less likely to occur in the first group meaning that it shows lower risk("protective") in risk of group 1 compared to 2. The odds ratio must be nonnegative if it is defined. It is undefined if p_2q_1 equals zero.

Confidence Intervals

The p values the authors often cite when reporting their results gives a sense of how likely the results reported are due to chance. However, p values do not allow us to make inferences about the precision of the estimates, which is extremely important in evaluating test characteristics. Reporting a range of plausible results, also known as confidence intervals, is more useful. Confidence intervals (CIs) are a measure of how precise an estimate is. The range or width of a confidence interval is primarily determined by two parameters; the number of observations in the study and how widely spread the data are (usually expressed as the standard deviation). The fewer the observations or the greater the data spread, the wider the confidence interval and the greater the uncertainty about the precision of the reported estimate.

Performance Measures

Once a data set is obtained, the problem arises of defining a measure for the quality of a particular prediction. Here, four different parameters are used to measure the performance of prediction method.

The predictive performance of a method is expressed by following four parameters:

1. Q *total*, the percentage of correctly classified residues, is defined as

$$Q\ total=\frac{p+n}{t}\times 100\%$$

where, p is the number of correctly classified beta-turn residues, n is the number of correctly classified non-beta-turn residues and t is the total number of residues in a protein. Q *total*, also known as 'prediction accuracy' may be defined simply as the total percentage of correct prediction. One difficulty with this measure is that it does not take into account disparities in the number of beta-turns (around 25%) and non-turns. Hence, it is possible to get a Q *total* score of about 75% by the trivial strategy of predicting all residues to be non-turn residues. Therefore, there is a risk of losing the information because of the dominance of non-turn residues. The Matthews Correlation Coefficient remedies this problem, which is defined as

2. MCC, the *Matthews Correlation Coefficient*, defined as

$$mcc=\frac{pn-ou}{\sqrt{(p+o)(p+n)(n+o)(n+u)}}$$

where, p is the number of correctly classified beta-turn residues, n is the number of correctly classified non-beta-turn residues, o is the number of non-beta-turn residues incorrectly classified as beta-turn residues and u is the number of beta-turn residues incorrectly classified as non-beta-turn residues. It is a measure that accounts for both over- and under-predictions.

3. Q *predicted*, defined as

$$Q\ predicted=\frac{p}{p+o}\times 100\%$$

Q *predicted* is the percentage of beta-turn predictions that are correct. Otherwise known as specificity, is the proportion of true negatives or the proportion of non-turn residues that have been correctly predicted as non turns.

4. Q *observed*, defined as

$$Q\ observed=\frac{p}{p+u}\times 100\%$$

Q *observed* is the percentage of observed beta-turns that are correctly predicted. Otherwise, known as sensitivity, is the proportion of true positives or the proportion of beta-turn residues that have been correctly predicted as beta-turns.

False Discovery Rates

Benjamini and Hochberg [13] introduced a multiple-hypothesis testing error measure called the false discovery rate FDR. This quantity is the expected proportion of false positive findings among all the rejected hypotheses. When testing multiple hypotheses, the situation becomes much more complicated. Now each test has type I and type II errors, and it becomes unclear how we should measure the overall error rate. Meanwhile, Multiple-hypothesis testing involves guarding against much more complicated errors than single-hypothesis testing. Whereas we typically control the type I error rate for a single-hypothesis test, a compound error rate is controlled for multiple-hypothesis tests. The first measure to be suggested was the family wise error rate FWER, which is the probability of making one or more type I errors among all the hypotheses. Instead of controlling the probability of a type I error at level α for each test, the overall FWER is controlled at level α. None-the-less, α is chosen so that FWER ≤ α and then a rejection region Γ is found that maintains level α FWER but also yields good power. We assume for simplicity that each test has the same rejection region, such as would be the case when using the p-values as the statistic. In many situations, FWER is much too strict, especially when the number of tests is large. Therefore, FDR is a more liberal, yet more powerful, quantity to control. In Storey [15], we introduced the positive false discovery rate pFDR. This is a modified, but arguably more appropriate, error measure to use. Benjamini and Hochberg [13] provided a sequential p-value method to control FDR. This is really what an FDR controlling p-value method accomplishes: using the observed data, it estimates the rejection region so that on average FDR ≤ α for some prechosen α. The product of a sequential p-value method is an estimate $\hat{k}$ that tells us to reject $p_{(1),}p_{(2),...,}p_{(\hat{k})}$, where $P_{(1)} \le P_{(2)} \le \ldots \le P_{(m)}$ are the ordered observed p-values. What can we say about $\hat{k}$. Is there any natural way to provide an error measure on this random variable? It is a false sense of security in multiple-hypothesis testing to think that we have a 100% guaranteed upper bound on the error. The reality is that this process involves estimation. The more variable the estimate of $\hat{k}$ is, the worse the procedure will work in practice. Therefore, the expected value may be that FDR ≤ α, but we do not know how reliable the methods are case by case. If point estimation only involved finding unbiased estimators, then the field would not be so successful. Therefore, the reliability of $\hat{k}$ case by case does matter even though it has not been explored. Another weakness of the current approach to false discovery rates is that the error rate is controlled for all values of m_0 (the number of true null hypotheses) simultaneously without using any information in the data about m_0. Surely there is information about m_0 in the observed p-values. Here, we use this information, which yields a less stringent procedure and more power, while maintaining strong control. Often, the power of the multiple hypothesis testing method decreases with increasing m. This should not be so, especially when the tests are independent. The larger the m, the more information we have about m0, and this should be used. In this paper, we propose a new approach to false discovery rates. We attempt to use more traditional and straightforward statistical ideas to control pFDR and FDR. Instead of fixing α and then estimating k (i.e. estimating the rejection region), we fix the rejection region and then estimate α. Fixing the rejection region may seem counter-intuitive in the context of traditional multiple-hypothesis testing. This can be seen to make sense in the context of false discovery rates. A natural objection to our proposed approach is that it does not offer 'control' of FDR. Actually, control is offered in the same sense as the former approach-our methodology provides a conservative bias in expectation. Moreover, since in taking this new approach we are in the more familiar point estimation situation, we can use the data to estimate m_0, obtain confidence intervals on pFDR and FDR, and gain flexibility in the definition of the error measure. We show that our proposed approach is more effective, flexible and powerful. The multiple hypothesis testing methods that we shall describe take advantage of more information in the data, and they are conceptually simpler.

The Positive False Discovery Rate and Fixed Rejection Regions

As mentioned in earlier, two error measures are commonly used in multiple-hypothesis testing: FWER and FDR. FWER is the traditional measure used; Benjamini and Hochberg [13] recently introduced FDR. Table 1 summarizes the various outcomes that occur when testing m hypotheses. V is the number of type I errors (or false positive results). Therefore, FWER is defined to be $\Pr(V \ge 1)$. Controlling this quantity offers a very strict error measure. In general, as the number of tests increases, the power decreases when controlling FWER. FDR is defined to be

$$FDR = E\left(\frac{V}{R}|R > 0\right)\Pr(R > 0)$$

i.e., the expected proportion of false positive findings among all rejected hypotheses times the probability of making at least one rejection. Benjamini and Hochberg [13] and Benjamini and Liu [14] provided sequential p-value methods to control this quantity. FDR offers a much less strict multiple-testing criterion over FWER and therefore leads to an increase in power. In Storey [15], we define a new false discovery rate, pFDR. Definition 1. $pFDR = E\left(\frac{V}{R}|R > 0\right)$. The term 'positive' has been added to reflect the fact that we are conditioning on the event that positive findings have occurred. When controlling FDR at level α, and positive findings have occurred, then FDR has really only been controlled at level α/Pr (R > 0). This can be quite dangerous, and it is not the case for pFDR. See Storey [15] for a thorough motivation of pFDR over FDR. Benjamini and Hochberg (1995) precisely define FDR to be expression (1) because this quantity can be controlled by a sequential p-value method. (Note, however, that weak control of FWER is implicitly embedded in this definition, i.e. FWER is controlled when all null hypotheses are true) pFDR is identically 1 when all null hypotheses are true ($m=m_0$), so the usual sequential p-value approach cannot be applied to this quantity. Therefore, to control pFDR, it must be estimated for a particular rejection region.

A sequential p-value method allows us to fix the error rate beforehand and to estimate the rejection region, which is what has traditionally been done in multiple-hypothesis testing. In the context of FWER this makes sense. Because FEWR measures the probability of making one or more type I error, which is essentially a '0–1' event, we can set the rate α *priori* at which this should occur. False discovery rates, in contrast, are more of an exploratory tool. For example, suppose that we are testing 1000 hypotheses and decide beforehand to control FDR at level 5%.Whether this was an appropriate choice largely depends on the number of hypotheses that are rejected. If 100 hypotheses are rejected, then clearly this was a good choice. If only two hypotheses are rejected, then clearly this was a less useful choice. Therefore fixing the rejection region beforehand can be more appropriate when using pFDR or FDR. For example, when performing many hypothesis tests, it can make sense to reject all p-values that are less than 0.05 or 0.01. Also, expert knowledge in a particular field may allow us to decide which rejection regions should be used. It will also be seen that this approach allows us to control pFDR, which we find to be a more appropriate error measure. Probably the most important reason for fixing the rejecting region is that it allows us to take a conceptually simpler approach to

Citation: Okeh UM, Okoro CN (2012) Evaluating Measures of Indicators of Diagnostic Test Performance: Fundamental Meanings and Formulars. J Biomet Biostat 3:132. doi:10.4172/2155-6180.1000132

complicated compound error measures such as pFDR and FDR. The q-value is the pFDR analogue of the p-value and allows the rejection regions to be determined by the observed p-values. This is more appropriate over either fixing the rejection region or fixing the error rate. But, by first fixing the rejection region in our approach, we can formulate the q-values quite easily.

Estimation and Inference for the Positive False Discovery Rate and False Discovery Rate

In this section, we apply the proposed approach to both pFDR and FDR.We first present a few simple facts about pFDR under independence to motivate our estimates. Suppose that we are testing m identical hypothesis tests H_1;H_2,…,Hm with independent statistics T_1; T_2,…,T_m. We let Hi = 0 when null hypothesis i is true, and Hi = 1 otherwise. We assume that the null Ti|Hi = 0 and the alternative Ti|Hi = 1 are identically distributed. We assume that the same rejection region is used for each test, which make the tests 'identical'. Finally, we assume that the Hi are independent Bernoulli random variables with Pr(Hi = 0) = π_0 and Pr(Hi = 1) = π_1. Let Γ be the common rejection region for each hypothesis test. The following is theorem 1 from Storey [15]. It allows us to write pFDR in a very simple form that does not depend on m. For this theorem to hold we must assume that the Hi are random; however, for large m this assumption makes little difference.

Theorem 1

Suppose that m identical hypothesis tests are performed with the independent statistics T_1,…,T_m and rejection region Γ. Also suppose that a null hypothesis is true with *a priori* probability π_0. Then

$$pFDR(\Gamma) = \frac{\pi_0 \Pr(T \in \Gamma \mid H = 0)}{\Pr(T \in \Gamma)}$$

$$= \Pr(H = 0 \mid T \in \Gamma)$$

$$where \Pr(T \in \Gamma) = \pi_0 \Pr(T \in \Gamma \mid H = 0) + \pi_1 \Pr(T \in \Gamma \mid H = 1)$$

This paper will be limited to the case where we reject on the basis of independent p-values. See Storey and Tibshirani [15] for a treatment of more general situations. For the remainder of the paper, instead of denoting rejection regions by the more abstract Γ, we denote them by γ, which refers to the interval [0; γ].

In terms of p-values we can write the result of theorem 1 as

$$pFDR(\gamma) = \frac{\pi_0 \Pr(T \le \gamma \mid H = 0)}{\Pr(T \le \gamma)} = \frac{\pi_0\,\gamma}{\Pr(T \le \gamma)}$$

where P is the random p-value resulting from any test. Under independence the p-values are exchangeable in that each comes from the null distribution (i.e. uniform [0,1]) with probability π_0 and from the alternative distribution with probability π_1. It is easiest to think about this in terms of simple *versus* simple hypothesis tests, but the theory also works for composite alternative hypotheses with a random effect [15]. Since π_0m of the p-values are expected to be null, then the largest p-values are most likely to come from the null, uniformly distributed p-values. Hence, a conservative estimate of π_0 is

Receiver Operating Characteristic Curve

ROC (Receiver Operating Characteristic) analysis is being used as a method for evaluation and comparison of classifiers [16]. The ROC gives complete description of classification accuracy as given by the area under the ROC curve. The ROC curve originates from signal detection theory [17]; the curve shows how the receiver operates the existence of signal in the presence of noise. The ROC curve plots the probability of detecting true signal (sensitivity) and false signal (1 – specificity) for an entire range of possible cut points. The sensitivity and specificity of a classifier also depend on the definition of the cut-off point for the probability of predicted classes. A ROC curve demonstrates the trade-off between true positive rate and false positive rate in binary classification problems. To draw a ROC curve, the true positive rate (TPR) and the false positive rate (FPR) are needed. TPR determines the performance of a classifier or a diagnostic test in classifying positive cases correctly among all positive samples available during the test. FPR, on the other hand, defines how many incorrect positive results, which are actually negative, there are among all negative samples available during the test. Because TPR is equivalent to sensitivity and FPR is equal to (1 –specificity), the ROC graph is sometimes called the sensitivity vs. (1 - specificity) plot. The area under the ROC curve has become a particularly important measure for evaluating classifiers' performance because it is the average sensitivity over all possible specificities [18]. The larger the area, the better the classifier performs. If the area is 1.0, the classifier achieves both 100% sensitivity and 100% specificity. If the area is 0.5, then we have 50% sensitivity and 50% specificity, which is no better than flipping a coin. This single criterion can be compared for measuring the performance of different classifiers analyzing a dataset [19,20]. After a classifier has been made, it is also useful to measure its calibration. Calibration evaluates the degree of correspondence between the estimated probabilities of a specific outcome resulting from a classifier and the outcomes predicted by domain experts. This can then be tested using goodness-of-fit statistics. This test examines the difference between the observed frequency and the expected frequency for groups of patients and can be used to determine if the classifier provides a good fit for the data. If the p-value is large, then the classifier is well calibrated and fits the data well. If the p-value is small, then the classifier is not well calibrated. There is a pair of diagnostic sensitivity and specificity values for every individual cut-off. To construct a ROC graph, we plot these pairs of values on the graph with the 1-specificity on the x-axis and sensitivity on the y-axis. Receiver operating characteristic curve analysis is often used to help determine the cut-off point to optimize sensitivity and specificity. An ROC curve is a graphical representation of the trade off between the false negative and false positive rates for every possible cut-off value [21]. Alternatively, the ROC curve is the representation of the trade off between sensitivity and specificity. In other words, the ability of a test using a specific analytic concentration, to discriminate disease from non-disease can be graphically portrayed by use of ROC curve analysis. A graph can be generated in which the sensitivity and specificity are determined for each data point obtained in the study. These are graphed with sensitivity of each data point on the y-axis and the corresponding 1-specificity for each data point on the x-axis. Precisely, we plot these pairs of values on the graph with the 1-specificity on the x-axis and sensitivity on the y-axis. (Note: the ratio of the y-axis/x-axis is the likelihood ratio positive or the graph of true positives and false positives respectively).For the ideal test, the plot would rise from 0 and go straight up to 1.00 and then a horizontal line along the 1.00 sensitivity line. This would be where there is no overlap in the data points and sensitivity and specificity would both be 100% in the left hand corner [21]. This rarely occurs and more commonly a curvilinear plot is observed. The greater the area under the curve, the more discriminatory (the ability of the test to correctly classify those with and without the disease) the test is, ideally, the area under a curve of 1.00 is a perfectly discriminatory test and a curve that follows the diagonal line in the graph has an area under the curve 0.5 which corresponds to the test being no better than flipping a coin [21]. The shape of a ROC curve and the area under the curve (AUC) helps us estimate how high is the discriminative power of a test. The closer the curve is located to upper-left hand corner and the larger the area under the curve, the better the test is at discriminating between diseased and non-diseased. The area under the curve can have any value between 0 and 1 and it is a good indicator of the goodness of the test. A perfect diagnostic test has an AUC 1.0. whereas a non-discrimination test has an area 0.5.The larger the area under the curve, the better the diagnostic test in discriminating those with and without disease [21]. Many statistical programs can generate a table of the values in the graph and calculate sensitivity, specificity, LP+,LP-,and proportion or percent correctly identified for each data point. Cut-off points are not necessarily chosen to optimize the number of patients correctly categorized. One can select different cut-off points to optimize sensitivity or specificity. For example, when a screening test is used to look for a serious disease that if missed could result in serious harm to the patient, the sensitivity of that test should be optimized. Conversely, in situations where therapy could be extremely harmful if given to a patient without the disease, one would choose a cut-off point that optimizes specificity. In general, when optimizing one test characteristic, the other gets worse and vice versa. For example, when improving sensitivity, specificity decreases and when improving specificity, sensitivity decreases. The area under the ROC curve can also be used statistically to compare the discriminating ability between two diagnostic tests [21]. We can say that the relationship between the area under the ROC curve(AUC) and diagnostic accuracy can be seen in the Table 2.1 below :

AUC is a global measure of diagnostic accuracy. It tells us nothing about individual parameters, such as sensitivity and specificity. Out of two tests with identical or similar AUC, one can have significantly higher sensitivity, whereas the other significantly higher specificity. Furthermore, data on AUC state nothing about predicative vales and about the contribution of the test in ruling-in and ruling-out a diagnosis. Global measures are there for general assessment and for comparison of two or more diagnostic tests. By the comparison of areas under the two ROC curves we can estimate which one of two tests is more suitable for distinguishing health from disease or any other two conditions of interest. It should be pointed that this comparison should not be based on visual nor intuitive evaluation [22]. For this purpose we use statistic tests which evaluate the statistical significance of estimated difference between two AUC, with previously defined level of statistical significance (P).

Youden's Index

Youden's index is one of the oldest measures for diagnostic accuracy [23]. It is also a global measure of a test performance, used for the evaluation of overall discriminative power of a diagnostic procedure and for comparison of this test with other tests. Youden's index is calculated by deducting 1 from the sum of test's sensitivity and specificity expressed not as percentage but as a part of a whole number: (sensitivity + specificity) – 1. For a test with poor diagnostic accuracy, Youden's index equals 0, and in a perfect test Youden's index equals 1. Youden's index is not sensitive for differences in the sensitivity and specificity of the test, which is its main disadvantage. Namely, a test with sensitivity 0,9 and specificity 0,4 has the same Youden's index (0,3) as a test with sensitivity 0,6 and specificity 0,7. It is absolutely clear that those tests are not of comparable diagnostic accuracy. If one is to assess the discriminative power of a test solely based on Youden's index it could be mistakenly concluded that these two tests are equally effective. Youden's index is not affected by the disease prevalence, but it is affected by the spectrum of the disease, as are also sensitivity specificity, likelihood ratios and DOR.

Area	Diagnostic Accuracy
0.9-1.0	Excellent
0.8-0.9	Very good
0.7-0.8	Good
0.6-0.7	Sufficient
0.5-0.6	Bad
< 0.5	Test not useful

Table 2.1: Relationship between the area under the ROC curve (AUC) and diagnostic accuracy.

	+Ve	–Ve	Total
+Ve	40	38	78
–Ve	5	237	242
Total	42	275	320

Table 2: Reference standard.

Design of Diagnostic Accuracy Studies

Measures of diagnostic accuracy are extremely sensitive to the design of the study. Studies suffering from some major methodological shortcomings can severely over- or under-estimate the indicators of test performance as well as they can severely limit the possible applicability of the results of the study. The effect of the design of the study to the bias and variation in the estimates of diagnostic accuracy can be quantified [24]. STARD initiative published in 2003 was a very important step toward the improvement of the quality of reporting of studies of diagnostic accuracy [25,26]. According to some authors, the quality of reporting of diagnostic accuracy studies did not significantly improve after the publication of the STARD statement [27,28], whereas some others hold that the overall quality of reporting has at least slightly improved [24], but there is still some room for potential improvement [29,30]. Editors of scientific journals are encouraged to include the STARD statement into the Journal Instructions to authors and to oblige their authors to use the checklist when reporting their studies on diagnostic accuracy. This way the quality of reporting could be significantly improved, providing the best possible evidence for health care providers, clinicians and laboratory professionals; to the best for the patient care.

Illustration of Some Statistics of Diagnostic Test Studies

To illustrate these, for our first example we shall use data from a study of diabetic eye tests [31]. This was a cross-sectional study in which diabetic patients being screening for eye problems were examined using direct opthalmoscopy (the test) and slit lamp stereoscopic biomicroscopy (the reference standard). A single sample of subjects all received both the diagnostic test and the reference standard test. The following table shows the results for all eye problems combined:

From this Table 2 we can calculate all diagnostic test statistics other than a ROC curve: sensitivity = 40/45 = 0.89 = 89%, specificity = 237/275 = 0.86 = 86%, LR (+ve test) = 0.89/(1 – 0.86) = 6.4, LR (-ve test) = 0.86/(1 – 0.89) = 7.8, OR = 40×237/(38×5) = 49.9, PPV = 40/78 = 51%, NPV = 237/242 = 98%. We shall now look at what these mean and how they were calculated. Sensitivity = the proportion of reference positive cases who are positive on the test =proportion of true cases that the test correctly identifies. Specificity = the proportion of reference negative cases who are negative on the test =proportion of true non-cases that the test correctly identifies. For eye disease in diabetics, there were 45 reference standard positive cases of whom 40 were positive on the test, 275 reference standard negative non-cases of whom 237

	+Ve	–Ve	Total
+Ve	120	10	130
–Ve	14	74	88
Total	134	84	218

Table 3: Reference standard.

were negative on the test. Sensitivity = 40/45 = 0.89 = 89%, Specificity = 237/275 = 0.86 = 86%. A good test will have high sensitivity and high specificity. We are looking for values exceeding 80%, preferably 90% or 95%. Odds = number of positives divided by number of negatives. Odds ratio (OR) = odds in one group divided by odds in another. For eye disease in diabetics: Odds test +ve for those reference +ve = 40/5 = 8.0, OR = (40/5)/(38/237) = 40×237/(38×5) = 49.9. As the test and the reference standard should have a strong positive relationship, we expect the odds ratio to be much greater than 1.0. The likelihood ratio (LR) for a positive test = sensitivity/(1 – specificity). We use this as follows. If we start with the probability that a subject has the disease, which is the prevalence of the disease, we can convert this to odds: odds = prevalence/(1 – prevalence). Then if we test a subject from a population with this prevalence, we can estimate the odds of having the disease if the test is positive: odds of disease if test positive = odds of disease × likelihood ratio. For eye disease in diabetics:

Likelihood ratio for a positive test = 0.89/(1 – 0.86) = 6.4. Suppose the prevalence of eye problem in the local diabetic population is 10% = 0.10. The odds of eye problems is 0.10/0.90 = 0.11. If a subject has a positive test, the odds of eye disease will be increased: odds of disease if test positive = 0.11 × 6.4 = 0.70. This corresponds to a probability of eye disease = 0.41. (Probability = odds/(1 + odds)). Similarly, the likelihood ratio for a negative test = specificity/(1 – sensitivity). As before, if we start with the probability that the subject does not have the disease = 1 – prevalence of disease and convert to odds = (1 – prevalence)/prevalence, we can look at the effect on the odds of not having the disease if the test is negative: odds of not disease if test negative = odds of not disease × likelihood ratio Likelihood ratio for a negative test = 0.86/(1 – 0.89) = 7.8. Suppose the prevalence of eye problem in the local diabetic population is 10% = 0.10. The odds of no eye problems is 0.90/0.10 = 9.0. If a subject has a negative test, the odds of no eye disease will be increased: odds of disease if test negative = 9.0 × 7.8 = 70.2. This corresponds to a probability of no eye disease = 0.986. The positive predictive value (PPV) is the proportion of test positives who are reference positive. The negative predictive value (NPV) is the proportion of test negatives who are reference negative. For eye disease in diabetics, there were 78 test positives of whom 40 were positive on the reference standard, 242 test negatives of whom 237 were negative on the reference standard. PPV = 40/78 = 51%., NPV = 237/242 = 98%. Hence if a subject is positive on the test, the probability that eye disease will be found using the reference standard is 51%. If a subject negative on the test, the probability that no eye disease will be found using the reference standard is 98%. For a receiver operating characteristic (ROC) curve, we need a different example. Sanchini et al. [32] looked at the early detection of bladder cancer using a test of elevated urine telomerase, an enzyme involved in cell proliferation. The reference standard was histologically confirmed bladder cancer. This was a case-control study conducted in 218 men: 84 healthy individuals and 134 patients at first diagnosis of histologically confirmed bladder cancer. Urine telomerase is a measurement taking a range of possible values rather the presence or absence of a sign. If we change the value of telomerase which we classify as elevated, this will change the sensitivity and specificity. We can do this and plot the sensitivity against the specificity to see how they vary together. For obscure historical reasons, it is usual to plot sensitivity against one minus specificity, also called the false positive rate. This is the ROC curve, a plot of sensitivity against 1 – specificity. (The name comes from telecommunications. As far as we are concerned, it is just a name.) This is the ROC curve of Sanchini et al. [32]: They have drawn two separate ROC curves, one for their whole sample and the other for men aged 75 years or older. Sensitivity increases as one minus specificity increase, i.e. as specificity decreases. We make our test more sensitive at the expense of making it less specific. We are looking for a compromise cut-off which will give both high sensitivity and high specificity. Sanchini et al. [32] chose 50 as being a reasonable compromise between a test which is sensitive, so finding most cases with the disease, and specific, so does not pick up a lot of people who do not have the disease. The diagnostic tests on a ROC curve do not have to determined by a continuous measurement, though they often are. All we need to plot the curve is more than one test. Sanchini et al. [32] also show a different, non-numerical test: urine cytology, not sensitive but fairly specific. For the detection of bladder cancer using as a test that urine telomerase > 50 against a reference standard of histologically confirmed bladder cancer, the 2 by 2 table is:

We can calculate most of the statistics as before: sensitivity = 120/134 = 0.90 = 90%, specificity = 74/84 = 0.88 = 88%. LR (+ve test) = 0.90/(1–0.88) = 7.5. LR (-ve test) = 0.88/(1–0.90) = 8.8. OR = 120×74/(10×14) = 63.4

However, the row totals would be meaningless and they are not shown in the table. This is because we took two separate groups of subjects. The row totals will depend on what ratio of cases to controls we used. They do not tell us anything about how many people would be test positive or test negative. As a result, PPV and NPV cannot be found in this study. We cannot estimate PPV and NPV in a case-control study. Their values depend on the prevalence of the disease in the population being tested. See Bland [33] for more information.

References

1. Irwig L, Bossuyt P, Glasziou P, Gatsonis C, Lijmer J (2002) Designing studies to ensure that estimates of test accuracy are transferable. BMJ 324: 669-671.
2. Raslich M, Markert R, Stutes S (2007) Selecting and interpreting diagnostic tests. Biochemia Medica 17: 139-270.
3. Sackett DL, Haynes RB, Guyatt GH, Tugwell P (1991) Clinical Epidemiology: a Basic Science for Clinical Medicine. Little Brown, Chicago.
4. Dohoo I, Martin W, Stryhn H (2003) Veterinary epidemiologic research. (Ist edn), Chariottetown, Prince Edward Island, Canada, pp: 706.
5. Staquett M, Rozencweig M, Lee YJ, Muggia FM (1981) Methodology for the assessment of new dichotomous diagnostic tests. J Chronic Dis 34: 599-610.
6. Sackett DL, Straus S, Richardson WS, Rosenberg W, Haynes RB (2000) Evidence-based medicine. How to practise and teach EBM. (2nd edn), Edinburgh: Churchill Livingstone 67-93.
7. Jaeschke R, Guyatt G, Lijmer J (2002) Diagnostic tests. In: Guyatt G, Rennie D, eds. Users' guides to the medical literature. Chicago: AMA Press, pp: 121-140.
8. Deeks JJ, Altman DG (2004) Diagnostic tests 4: likelihood ratios. BMJ 329: 168-169.
9. Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PMM (2003) The diagnostic odds ratio: a single indicator of test performance. J Clin Epidemiol 56: 1129-1135.
10. Cornfield J (1964) A Method for Estimating Comparative Rates from Clinical Data. Applications to Cancer of the Lung, Breast, and Cervix. Journal of the National Cancer Institute 11: 1269-1275.
11. Mosteller F (1968) Association and Estimation in Contingency Tables. Journal of the American Statistical Association 63: 1–28.
12. Edwards AWF (1963) The measure of association in a 2x2 table. Journal of the Royal Statistical Society Series A 126: 109-114.

13. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. JR Statist Soc B 57: 289-300.

14. Benjamini Y, Liu W (1999) A step-down multiple-hypothesis procedure that controls the false discovery rate under independence. J Statist Plangg Inf 82: 163-170.

15. Storey JD (2001) The positive False Discovery Rate: a Bayesian interpretation and the q-value. The Annals of Statistics 31: 2013-2035.

16. Ferri C, Flach P, Hernandez-Orallo J (2002) Learning Decision Trees Using the Area under the ROC Curve. Nineteenth International Conference on Machine Learning (ICML 2002), Morgan Kaufmann, pp: 46-139.

17. Hosmer DW, Lemeshow S (2000) Applied Logistic Regression.(2nd edn), Wiley, Inc, New York.

18. Bradley AP (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition 30: 1145-1159.

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

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

21. Zweig M, Campbell G (1993) Receiver operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin Chem 39: 561.

22. Obuchowski NA, Lieber ML, Wians FH Jr (2004) ROC curves in clinical chemistry: Uses, misuses, and possible solutions. Clin Chem 50: 1118-1125.

23. Youden WJ (1950) Index for rating diagnostic tests. Cancer 3: 32-35.

24. Rutjes AWS, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, et al. (2006) Evidence of bias and variation in diagnostic accuracy studies. CMAJ 174: 469-476.

25. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, et al. (2003a) Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Clin Chem 49:1-6.

26. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, et al. (2003b) The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Clin Chem 49: 7-18.

27. Wilczynski NL (2008) Quality of reporting of diagnostic accuracy studies: no change since STARD statement publication--before-and-after study. Radiology 248: 817-823.

28. Bossuyt PMM (2008) STARD statement: still room for improvement in the reporting of diagnostic accuracy studies. Radiology 248: 713-714.

29. Bossuyt PMM (2004) The quality of reporting in diagnostic test research: getting better, still not optimal. Clin Chem 50: 465-466.

30. Bossuyt PMM (2006) Clinical evaluation of medical tests: still a long road to go. Biochemia Medica 16: 89-228.

31. Harding SP, Broadbent DM, Neoh C, White MC, Vora J (1995) Sensitivity and specificity of photography and direct ophthalmoscopy in screening for sight threatening eye disease: the Liverpool diabetic eye study. BMJ 311: 1131-1135.

32. Sanchini MA, Gunelli R, Nanni O, Bravaccini S, Fabbri C, et al. (2005) Relevance of urine telomerase in the diagnosis of bladder cancer. JAMA 294: 2052-2056.

33. Bland JM (2004) Interpretation of diagnostic tests.

34. Giard RWM, Hermans J (1996) The diagnostic information of tests for the detection of cancer: the usefulness of the likelihood ratio concept. Eur J cancer 32: 2042-2048.