Course Notes. Evaluation of Diagnostic Tests: Receiver Operating Characteristic Curves

Evaluarea testelor diagnostice: curbele ROC

Cristian Baicus[1,2*], Adriana Hristea[2,3], Anda Baicus[3,4]

1. Carol Davila University of Medicine, Colentina University Hospital, Department of Internal Medicine, Bucharest, Romania
2Clinical Research Unit RECIF (Réseau d'Epidémiologie Clinique International Francophone), Bucharest, Romania
3 "Prof. Dr. Matei Bals" National Institute for Infectious Diseases, Bucharest, Romania
4 "I. Cantacuzino" National Institute of Research and Development in Microbiology-Immunology Bucharest, Romania

Abstract

Receiver operator characteristic (ROC) curves are used in order to assess the accuracy of diagnostic tests whose results are continuous numeric variables. This curve is a graph of the sensitivity (or true positive rate) on the Y-axis as a function of 1-specificity (the false positive rate) on the Y-axis. ROC curves have multiple utilizations: 1. The comparison of more tests for the same disease (bigger the area under the ROC curve (AUROC), better the test; an AUROC of 1 means a perfect test, while an AUROC of 0.5 means a useless test; 2. The choice of a cut-off point (in case of a test with a big AUROC, one can choose the closest point to the upper left corner of the graph, in order to have both a good sensitivity and a good specificity); 3. It shows how, for the same diagnostic test, there is a negotiation between sensitivity and specificity so that, for a cut-off with a very good sensitivity it will be a weak specificity and the reverse.

Rezumat

Pentru evaluarea acurateții testelor diagnostice cu rezultate numerice, se folosesc curbele ROC, care reprezintă graficul sensibilității (pe axa X) in funcție de rata fals pozitivilor (pe axa Y). Aceste curbe au mai multe utilizări: 1. compararea mai multor teste pentru aceeaşi boală (cu cât aria de sub curbă este mai mare, cu atât este testul mai acurat; o arie de 1 reprezintă un test perfect, iar o arie de 0,5 reprezintă un test inutil, care nu are nici un aport in diferențierea celor care au boala de cei sănătoşi); 2. stabilirea valorilor de prag în funcție de care rezultatul testului este pozitiv sau negativ (pentru testele cu o arie de sub curbă mare, pentru a avea simultan o sensibilitate şi specificitate bune, alegem ca valoare de prag punctul situat cel mai aproape de colțul din stânga-sus al graficului); 3. ne demonstrează cum, pentru acelaşi test diagnostic, există o negociere între sensibilitate şi specificitate, astfel incât, dacă alegem o valoare de prag pentru o foarte bună sensibilitate, vom avea o specificitate slabă şi invers.

[*]**Corresponding author:** Dr. Cristian Baicus, Colentina University Hospital 3rd Department of Internal Medicine, Soseaua Stefan cel Mare 19-21, Sector 2 020125 Bucharest, Romania
Tel: +40 788302355 Fax: +40 213180657 E-mail: cbaicus@clicknet.ro

Introduction

In the process of evaluation of diagnostic tests, the used parameters are sensitivity, specificity, positive and negative predictive values and likelihood ratios, and they are always calculated reporting to a reference test (gold standard), considered perfect.

When the results of the diagnostic tests are dichotomal (test positive or negative), sensitivity, specificity and predictive values are computed in a 2 by 2 contingency table, while the likelihood ratios are computed using sensitivity and specificity (*Table 1*).

Designing receiver operating characteristic (ROC) curves

However, the results of many clinical tests are quantitative and are provided on a continuous scale. In order to decide if a test is positive or negative, a cutoff point has to be set, and according to this cutoff point, one can compute all the parameters mentioned above.

With such tests whose results are continuous variables, several values of sensitivity and specificity are possible, depending on the cutoff point chosen to define a positive test. This trade-off between sensitivity and specificity can be displayed using a graphic technique originally developed in electronics: receiver operating characteristic (ROC) curves. The investigator selects several cutoff points and determines the sensitivity and specificity at each point and then graphs the sensitivity (or true positive rate) on the Y-axis as a function of 1-specificity (the false positive rate) on the Y-axis (*Figure 1)*.

As one can see on the graph, if specificity is figured on the top of the graph, paralleling the X-axis, perfect (100%) specificity will be at the intersection with Y-axis (at the same point with perfect=100% sensitivity). This means that an ideal test is one that reaches the upper left corner of the graph (100% true positives and no false positive). A worthless test follows the diagonal from the lower left to the upper right corner: at any cutoff, the true positive rate is the same as the false positive rate.

Table 1: 2x2 contingency table for the evaluation of a diagnostic test

		DISEASE		TOTAL
		PRESENT	ABSENT	
DIAGNOSTIC TEST	POSITIVE	*a*	*b*	*a+b*
	NEGATIVE	*c*	*d*	*c+d*
		a+c	*b+d*	*a+b+c+d*

(a=true positive; b=false positive; c=false negative; d=true negative)
Sensitivity (Sn) = a/(a+c)
Specificity (Sp) = d/(b+d)
Likelihood ratio for a positive result of the test (LR+) = **sensibivity**/(1-**specificity**)
Likelihood ratio for a negative result of the test (LR-) = **specificity**/(1-**sensitivity)**
Pretest odds = **prevalence/**(1-**prevalence)**
Posttest odds = pretest **odds x likelihood ratio**
Posttest probability = posttest **odds**/(posttest **odds** +1)
Pretest probability = prevalence = (a+c)/(a+b+c+d)
Predictive value, positive (PPV) = a/(a+b)
Predictive value, negative (NPV) = d/(c+d)

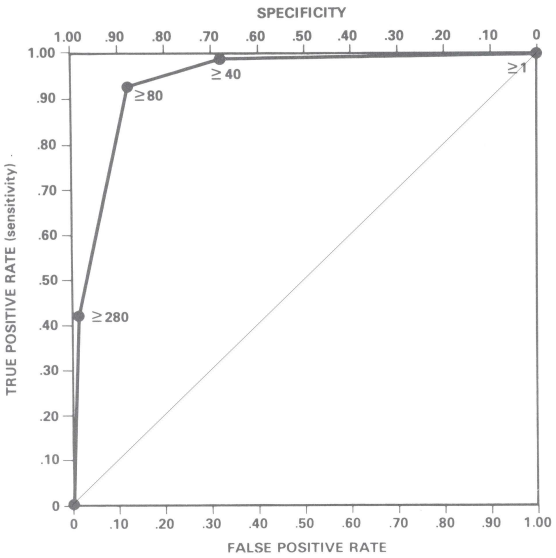

Figure 1. ROC curve of CK in acute myocardial infarction

Comparing more tests for the same disease

The area under the ROC curve (AUROC), which thus ranges from 0.5 for a useless test to 1.0 for a perfect test, is a useful summary of the overall accuracy of a test and can be used to compare the accuracy or two or more tests. As seen in *Figure 2*, serum iron has the lowest value in the diagnosis of anemia due to iron deficiency, because its curve is the closest to the diagonal line (corresponding to an AUROC of 0.5), while ferritin is the best test, because its curve is the closest to the upper left corner (and the furthest to the diagonal reference line). The AUROCs are 0.597 for serum iron, 0.757 for the percentage iron saturation, and 0.824 for ferritin.

This property is useful when you are trying to decide which of two or more competing tests for the same target disorder is the better one.

On the other way, for any AUROC value resulting from a diagnostic study, the 95% confidence intervals is computed; for a test to be useful, this confidence interval must not contain the value of 0.5, which is the area of a useless test.

Choosing a cut-off point

When a test is accurate, and thus the ROC curve is close to the upper left corner (and AUROC is close to 1), a cut-off point for both a good sensitivity and specificity can be choosen, and this is the closest point to the upper left corner. For example, in *Figure 1*, if we choose 80 U/L of creatinkinase (CK) as a cut-off point, the test has a sensitivity (SN) = 0.93 (93%) and a specificity (Sp) = 0.89 (89%) for acute miocardial infarction. For a cut-off value of 40 U/L, the test becomes more sensitive (99%), but less specific (68%), and this means that, if negative, we are sure the patient has not miocardial infarction, but if positive, we cannot say the patient has the disease. For a cut-off value of 280 U/L, the test becomes, on the contrary, more specific (99%), but much less sensitive (43%)

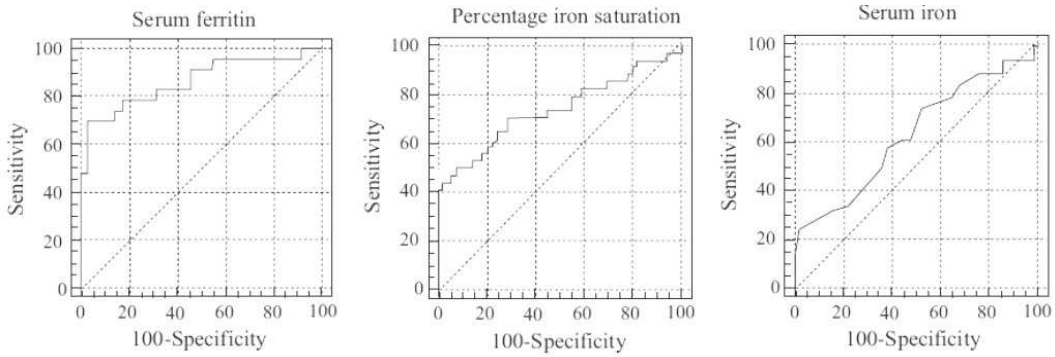

Figure 2. Comparison of more tests for the same disease: ROC curves of serum iron, transferin saturation and ferritin in iron deficiency anemia (1)

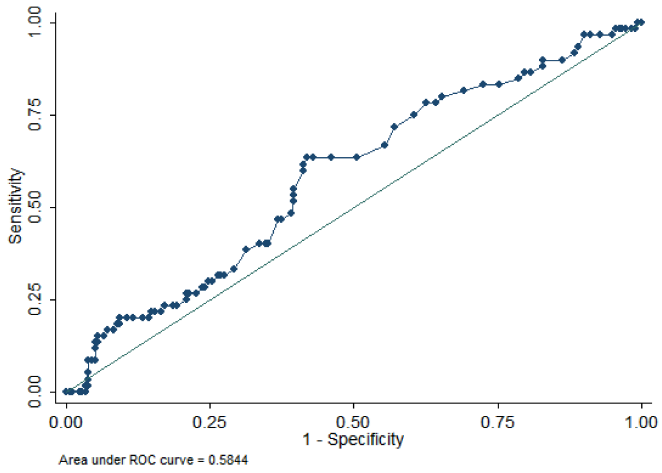

Figure 3. ROC curve of RDW for the diagnosis of cancer in involuntary weight loss

and this means that, if positive, we are sure the patient has a miocardial infarction, while if negative, we cannot exclude a miocardial infarction at all! This example demonstrates the fact that, for a given test, any increase in sensitivity will be accompanied by a decrease in specificity, and vice versa.

In *Figure 3* is plotted the ROC curve of red cell distribution width (RDW) for the diagnosis of cancer in patients with involuntary weight loss (2). Because the AUC for RDW is closer to 0.500 (a useless test) than to 1.00 (a perfect test), one cannot choose a cut-off value for which the test would have both good sensitivity and specificity, because such a cut-off does not exist, as the curve is far from the upper left corner. In this case, it would be wiser to use RDW as a multilevel test which has, for a value lesser than 12.7%, a good sensitivity (94%), and for a value higher than 18.4%, a good specificity (94%). However, only a few patients (41 of 253, 16%) could benefit of one of these extreme values because most of them, with cancer or not, had RDW between 12.7 and 18.4%.

In cases where you must choose a single cutoff point for an interval test, it is best to do it based on the clinical implications of false positive and false negative results. When it is important not to miss a diagnosis, you need tests that are highly sensitive. On the other hand, before subjecting patients to dangerous or painful interventions, you need tests that are highly specific (3).

Although useful, the AUROC does not provide information about how patients are misclassified. For example, a test with a high sensitivity but relatively low specificity may be useful for case-finding or screening, and one with low sensitivity and high specificity may be appropriate to „rule-in" a disease. Nevertheless, both may have a mediocre area under the curve (4). Thus, the physicians should also utilize another measure of performance, such as likelihood ratios or predictive values.

As showed above, the ROC curve shows how severe the trade-off between sensitivity and specificity is for a test and can be used to help decide where the best cutoff point should be.

Obviously, tests that are both sensitive and specific are highly sought after and can be of enormous value. However, practitioners must frequently work with tests that are not highly sensitive and specific. In these instances, they must use other means of circumventing the trade-off between sensitivity and specificity. The common way is to use the results of several tests together.

References

1. Ong KH, Tan HL, Lai HC, Kuperan P. Accuracy of various iron parameters in the prediction of iron deficiency in an acute care hospital. Ann Acad Med Singapore.2005;34:437-40
2. Baicus C, Ionescu R, Tanasescu C. Does this patient have cancer? The assessment of age, anemia, and erythrocyte sedimentation rate in cancer as a cause of weight loss. A retrospective study based on a secondary care university hospital in Romania. Eur J Intern Med. 2006 Jan;17(1):28-31.
3. Newman TB, Browner WS, Cummings SR. Designing studies of medical tests. In: Hulley SB, Cummings SR, Browner WS, Grady D, Hearst N, Newman TB. Designing clinical research. 2nd Edition, Lippincot Williams & Wilkins, Philadelphia, 2001, p. 175-194.
4. Katz MH. Study design and statistical analysis. A practical guide for clinicians. Cambridge University Press New York, 2006, p. 146-148.