II. Process Studies (Diagnostic Test)

Juan O. Talavera, Niels H. Wacher-Rodarte, Rodolfo Rivas-Ruiz

The purpose of a diagnostic test is to establish the presence of health or disease, it can even graduate the degree of illness. Diagnostic tests are usually assessed mathematically. Thus, sensitivity and specificity are estimated once the existence or not of disease is known; in clinical practice, the course of action is often reverse: from positivity or negativity of a test for the presence or not of the disease and, therefore, positive and negative predictive values are used. Mathematical strategies allow for an observation to be quantified, but clinical judgement is required in order to establish the quality of that observation; in consequence, some characteristics have to be considered: *a)* selection under the same criteria for cases and witnesses; *b)* inclusion of the entire spectrum of severity of the disease (trying that all the strata include an important number of subjects); *c)* the interpretation of the gold standard and the test under study must be blinded and done by experts; *d)* the interpretation of the results must show the applicability of the test in everyday practice; *e)* reproducibility of the test must be proven. It is important not to forget that, usually, only a patient is seen at a time; therefore, full knowledge of the diagnostic test performance is essential, as well as considering the clinical aspects for its correct application.

Key words
research
research projects
diagnostic techniques and procedures

This article was originally published in Rev Med Inst Med Seguro Soc 2011; 49 (2): 163-170 and it has been reviewed for this issue.

Introduction

Part 1 of this series [Rev Med Inst Seguro Soc 2011; 49(1):53-58] mentioned the different approaches for addressing clinical problems: *architectural approach*, based on the natural phenomenon; *methodological approach*, based on the hierarchy of the information; *clinical approach*, based on the aims of medical practice. Methodological approach key features were analyzed in detail, and integration studies were also mentioned.

However, in clinical practice, questions use to be related with the need to establish a diagnostic or ascribe causality either through a prognostic study, a treatment, or by trying to identify whatever provoqued a certain disorder or disease. This is where the architectural approach fits together with the objective-based approach.

Among the process studies, according to the architectural approach there is the diagnostic testing (objective-based approach). Additionally, causality studies include the prognostic, treatment and risk factors or causative agent studies (objective-based approach). In this article, we describe the most commonly used tools in diagnostic testing.

In clinical practice, a diagnostic test aims to identify the health or disease status of the subject under study. Frequently, in the presence of a disease, it allows for the severity of the condition to be established; for example: in a patient with sudden neurological deficit, tomography allows for the diagnosis to be defined (ischemic stroke), whereas if the diagnosis is already available, tomography allows for the extent of the lesion to be known.

The use of mathematics during the diagnostic process has the purpose of estimating the degree of efficacy and certainty of the tests in clinical practice. Below, the main features of every diagnostic test, using both clinical data and laboratory and imaging findings, are described.

Characteristics of a Diagnostic Test

The way to assess the efficacy of a diagnostic test depends on the type of data (variable) to be used. Therefore, it is important to identify the type of variable. Basic variables are those that we know as *qualitative of the nominal* or *dichotomic type*, and they refer to those for which we only notice its presence or for which only two options exist (e.g., nationality, presence or not of disease, male or female). *Ordinal qualitative* variables are those in which it can be identified only the place occupied in the group by the evaluated characteristics, but we do not know the size of the difference between

each other (e.g., the degree of severity of a disease —mild, moderate or serious—, or the intensity of a clinical piece of information identified by a cross mark, where, even when + is acknowledged to be lower than ++ and, consequently, lower than +++, ++ can not be stated as being double to +). And, finally, *quantitative variables* are those in which the distance between two levels of intensity is known; and in this variables the distance between two units is always equidistant. They are known as *discrete* or *discontinuous* when they can not be fractionated (e.g., how many children has a family [0, 1, 2, 3]), and *continuous* when fractions can be identified between one value and another (e.g., 52.0 kg, 52.2 kg or 52.250 kg weight).

Sensitivity and specificity are distinctive characteristics of every diagnostic test and indicate their efficacy. *Sensitivity* refers to the proportion of diseased individuals with a positive test. *Specificity* refers to the proportion of non-diseased individuals with a negative test.

The calculation of sensitivity and specificity uses *nominal* or *dichotomic* data and it is based on the use of a 2 × 2 table, in which the tested data is contrasted against the final diagnosis obtained by means of an ideal parameter named *gold standard*, which represents the test with the highest reliability for demonstrating a disease, e.g., histopathological results (testicular seminoma), surgical findings (cholecystitis), imaging studies interpretation (stroke by tomography or magnetic resonance imaging), interventional imaging studies (type of congenital heart disease by cardiac catheterization) or laboratory findings (renal failure by creatinine clearance).

Figure 1 shows the calculation of sensitivity and specificity of neck stiffness for the diagnosis of subarachnoid hemorrhage in patients with sudden onset neurological deficit, likely of vascular cause. A sensitivity of 59 % with a specificity of 94 % is observed, which means that 59 % of the patients with subarachnoid hemorrhage may show neck stiffness and among those without subarachnoid hemorrhage, 94 % do not have neck stiffness.

Sensitivity and specificity calculations are directed from the presence or absence of a particular disease, towards the probability of experiencing or not certain data. However, in clinical practice, the approach is often in the reverse direction: it goes from a positive or negative test result to the likelyhood of having or not a specific disease. This type of orientation corresponds to what we know as predictive values. The *positive predictive* value represents the probability that a patient with a certain positive test (sign, symptom, laboratory or imaging result or some index) has of suffering a particular disease; the *negative predictive value* is the probability that a patient, with a certain negative test, has of being free from a particular disease.

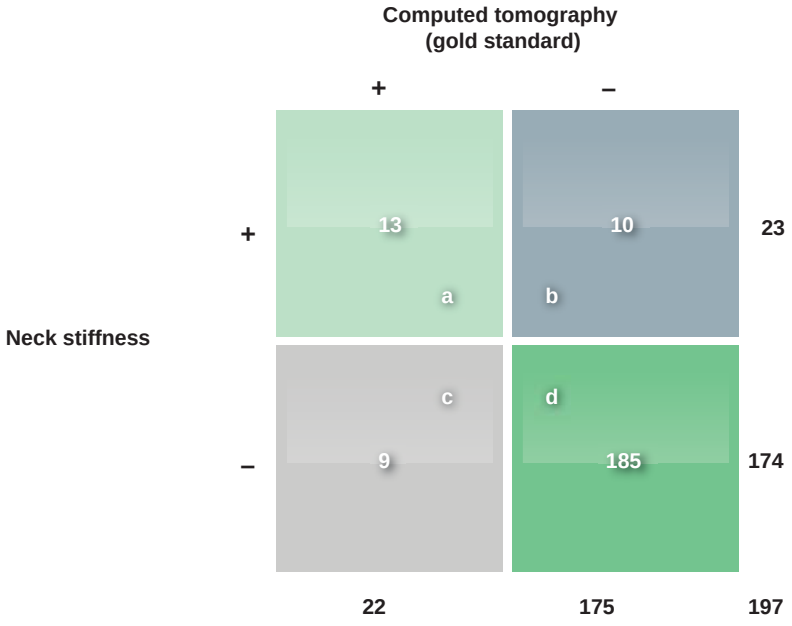


Figure 1 Sensitivity and specificity estimation of neck stiffness in the diagnosis of subarachnoid hemorrhage

Sensitivity	a/a + c = 0.59 (59 %)	Specificity	d/b + d = 0.94 (94 %)
False positives	b/b + d = 0.6 (6 %)	False negatives	c/a + c = 0.41 (41 %)
Positive predictive value	a/a + b = 0.57 (57 %)	Diagnostic certainty	d/c + d = 0.95 (95 %)
Prevalence	a + c /a + b + c + d = 0.11 (11 %)	Certeza diagnóstica	a + d/a + b + c + d = 90 (90 %)

Figure 1 shows a positive predictive value of 57 % and a negative predictive value of 95 %; this means that among the patients with clinical symptoms of stroke, a subject with neck stiffness has a 57 % probability of suffering from subarachnoid hemorrhage, whereas a patient without neck stiffness has a 95 % probability of not having subarachnoid hemorrhage.

While sensitivity and specificity values are considered to be constant, which is not true as we will explain later, predictive values are affected by disease prevalence. For example, in Figure 2, where the disease prevalence increased only from 11 to 56 %, maintaining the proportion of diseased subjects with positive and negative tests, sensitivity and specificity are shown to be preserved, whereas predictive values change: the positive predictive value is 93 % and the negative predictive value is 65 %. Thus, a prevalence increase causes an increase in the positive predictive value, with a decrease in the negative predictive value (a positive test in a population with high prevalence of the disease practicaly establishes the diagnosis; a negative test, however, does not rule it out); conversely, a decrease in prevalence produces an increase in the negative predictive value and a decrease in the positive predictive value (a negative test in a population with low prevalence of the disease almost rules the disease out).

If prevalence of the disease in the population from which predictive values of the diagnostic test were obtained is different from the prevalence of the disease in our population, these predictive values cannot be used. However, Bayes' theorem allows for predictive values to be estimated by using the sensitivity and specificity of the test, as well as the prevalence of the entity under study in our population. Table I shows how the increase in prevalence from 11 to 56 % produces a 57 to 94 % increase in the positive predictive value. This example shows clearly how a positive test in a population with low prevalence (11 %) has an approximate probability of 50 % for diagnosing the disease, whereas with a high prevalence (56 %), it practically establishes the diagnosis.

Another practical strategy for estimating the probability of the disease in case of a positive test, but at different prevalence values, is the use of Fagan's nomogram and the likelihood ratio (LR). The positive LR (PLR) is obtained from the ratio sensitivity/1-specificity. In turn, the negative LR (NLR) is obtained from the ratio 1-sensitivity/specificity. Fagan's nomogram is divided in three parts. In the first column appears the pre-test possibility (prevalence). In the middle, there are the values of the LR and in the last column, the post-test probability. The post-test probability for a PLR refers to the probability of obtaining a positive result when the test is positive and it corresponds to the PPV; the post-test probability for an NLR refers to the probability of obtaining a positive result when the test is negative, which is equivalent to 1-NPV. Examples for a prevalence of 11 and 56 % are shown in Figure 3.

It was mentioned previously that the sensitivity and specificity of a test are not dependent on

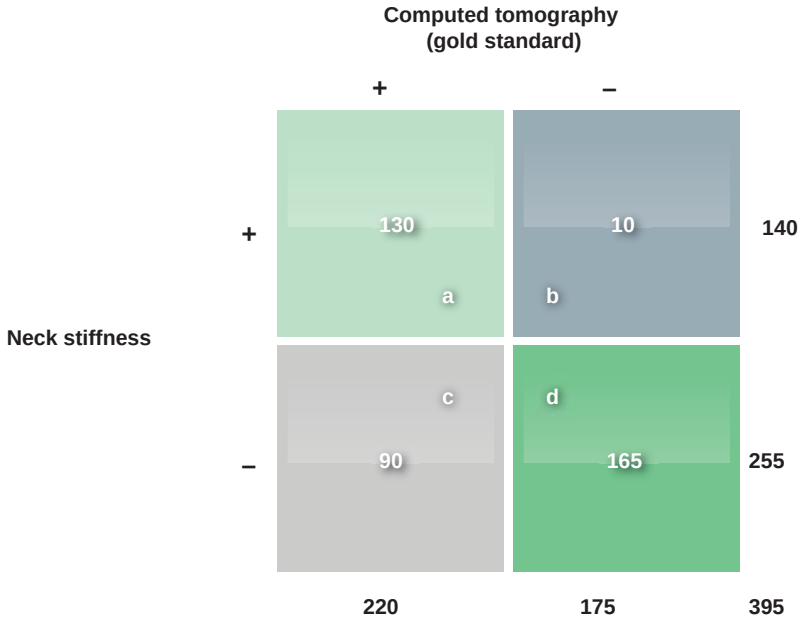


Figure 2 Modification of neck stiffness predictive values in the diagnosis of subarachnoid hemorrhage with the change in prevalence

Sensitivity = 59 %	Specificity = 94 %	
Positive predictive value = 93 %	Negative predictive value = 65 %	Prevalence = 56 %

the prevalence of the disease; however, the values vary according to the predominant disease severity degree in the group under study. For example, diagnosing lung cancer at an advanced stage with a chest x-ray is simple and it will rarely go unnoticed, i.e., false negatives will rarely exist and sensitivity will be high; however, it will be hardly detected if we try to diagnose it in asymptomatic individuals, at an early stage, which will provoke a high percentage of false negatives and low sensitivity. Therefore, considering that the sensitivity obtained from a test in a population is applicable to other population implies that the distribution of disease severity is the same in both samples, since if in the first one the proportion of subjects in advanced stages is predominant, sensitivity will be high, and if in the second prevails an early stage, sensitivity will be low. Having the same inclusion criteria between different studies of different populations does not guarantee that the distribution of subjects will preserve a similar proportion of subjects at every stage of the disease and, consequently, sensitivity may be different.

Use of Ordinal and Quantitative Data

Unlike nominal data, when the test under study corresponds to ordinal or quantitative data (with more than one cut-off point), a ROC (receiver operator characteristic) curve has to be plotted, which enables to determine in which of the cut-off points the highest diagnostic certainty is obtained.

Figure 4 shows the different value ranges of creatine phosphokinase in cerebrospinal fluid expressed in µU/mL, with their respective frequencies, and the calculation of sensitivity and specificity is outlined according to the different cut-off points by elaborating 2 × 2 tables. In these tables, intervals are constructed with the different values of the test under study and tabulated in two columns; the first shows the frequencies of subjects with the disease in each of the intervals and the second shows the frequency of subjects without the disease within the same intervals. The most altered values appear above (first intervals) and the less altered below. The cumulative percentaje is calculated upwards and downwards of each cut-off point, in both columns. In the column of diseased subjects, sensitivity is estimated from the cut-off point upwards, and in the column of controls, the percentage of false positives (1-specificity).

The results are plotted with the sensitivity values and the percentage of false positives: sensitivity values on the ordinate axis (*Y*), and the ratio of false positives (1-specificity) on the abscissa axis (*X*); a specificity value of 90 % corresponds to 10 % of false positives (Figure 5). The best cut-off point corresponds within the ROC curve to the closest point to the left superior angle of the curve, or to the point within the table that contains the lowest $b + c$ value (values that belong to the sum of false positives and false negatives) or the highest value for $a + d$ (values that belong to the sum of true positives and true negatives). In this case, the cut-off point is ≥ 16 µU/mL, which allows for 79.6 % of patients to be correctly classified as diseased or healthy, with a sensitivity of 61.5 % and a specificity of 96.5 %. However, according to the use given to the test, more than one point can be selected: where sensitivity or specificity is favored (higher negative or positive predictive value).

There are cases in which not only the test under study contains more than two strata, but even the gold standard. In these cases the percentage of success and error can be estimated. Figure 6 compares clinical diagnosis of pulmonary embolism considering the diagnosis by ventilation/perfusion scan as the gold

Table I Bayes' theorem

$$p\,(E+/P+) = \frac{p\,(P+/E+)\ \mathrm{p}\,(E+)}{p\,(P+/E+)\,p\,(E+) + p\,(P+/E-)\,p\,(E-)}$$

p (E+/P+) =	*a posteriori* probability of having a certain disease in case of a positive test; corresponds to the *positive predictive value* (PPV).
p (P+/E+) =	probability of a positive test result when the patient has the disease; corresponds to *sensitivity*.
p (E+) =	*a priori* probability of having the disease according to the population that the subject belongs to; corresponds to *prevalence*.
p (P+/E–) =	*probability of a positive test result when the patient does not have the disease; equivalent to false positives or 1-specificity.*
p (E–) =	*a priori* probability of not having the disease and corresponds to 1-prevalence. [1 – p (E+)].

Prevalence	11 %	56 %
Sensibility	59 %	59 %
Specificity	94 %	94 %
PPV	57 %	94 %
NPV	95 %	64 %

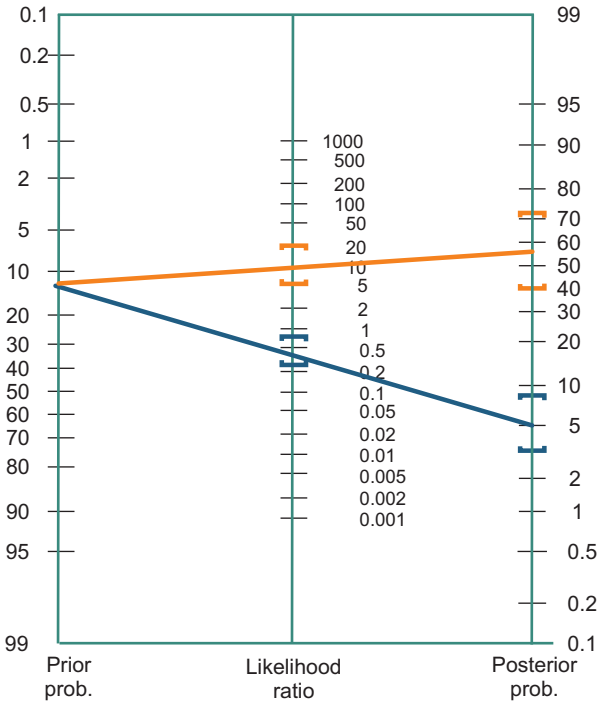
The negative predictive value is estimated in the same way reversing the signs of the formula [e.g.: p (E+ /P+) changes to p (E-/P-)]

standard; the percentage of accuracy corresponds to the cells where both clinical diagnosis and the gold standard match, i.e. in cells *a, e, i* (40 + 90 + 70), with this being 66.66 %, and our percentage of errors overestimating the diagnosis corresponds to cells *b, c, f* (30 + 20 + 10), with this being 20 %; finally, the percentage of error underestimating the diagnosis is comprised by cells *d, g, h* (7 + 30 + 3), with this being 13.33 %. However, there is the possibility of wanting to handle the outcome only with two possibilities; in this case, the scans with low and moderate probability could be grouped and talk about a scan with high probability of pulmonary embolism or without high probability, or grouping those with high and intermediate probability and leaving those with low probability in a single group. This same procedure can be performed with the clinical scale, so that by having only four cells, the traditional usefulness estimators of a diagnostic test can be used, or preserving the three strata of our test under study and calculate a ROC curve.

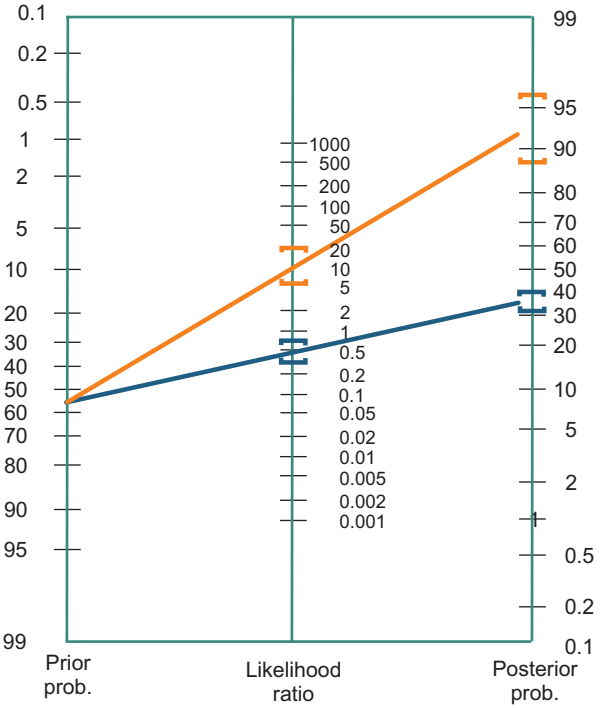
Diagnostic Test Applications

It should remain clear that the application of a test may have different purposes:

1. If a screening test is wanted, a high sensitivity test should be used, even if it has low specificity (e.g., test strips to measure blood glucose, to search for suspected diabetes mellitus).
2. If ruling out a given disease is wanted, a test with high sensitivity and, if possible, high specifity is used (high negative predictive value, e.g., ELISA for HIV), since, although when positive it is not diagnostic, when negative it does rule it out.
3. If we want to confirm a diagnosis in a patient suspected of having a certain disease, a test with high specifity and, if possible, high sensitivity is used (high positive predictive value, e.g., Western-Blot for HIV), since, although when negative it does not always rule the disease out, if positive, it establishes the diagnosis.



Prevalence = 0.11 (11 %)
Sensitivity = 0.59 (59 %)
Specificity = 0.94 (94 %)
Positive predictive value = 0.57 (57 %)
Negative predictive value = 0.95 (95 %)
Positive likelihood ratio = (a/a + c)/1-(d/b + d)= 9.83
Negative likelihood ratio = 1-(a/a + c)/(d/b + d)= 0.44
Post-test probability for PLR = 57 %
Post-test probability for NLR = 5 %



Prevalence = 56 %
Sensitivity = 59 %
Specificity = 94 %
Positive predictive value = 93 %
Negative predictive value = 65 %
Positive likelihood ratio = 9.83
Negative likelihood ratio = 0.44
Post-test probability for PLR = 93 %
Post-test probability for NLR = 35 %

Figure 3 Use of Fagan's nomogram and likelihood ratios

Creatine phosphokinase (μU/mL)	Damage		≥ 76	≥ 61	≥ 46	≥ 31	≥ 16	≥ 0
	Organic	Metabolic						
≥ 76	2	0	2	0				
61-75	1	0	25	27	3	0		
46-60	2	0			24	27	5	0
31-45	6	0					22	27
16-30	6	1						
≤ 15	10	26						

					16	27	17	1
							10	26	27	27
									0	0

a b
c d

Sensitivity	(a/a + c)	%	7	11	19	41	63	100
Specificity	(d/b + d)	%	100	100	100	100	96	0
False positives		%	0	0	0	0	4	100
Error			25	24	22	16	11	27

Figure 4 Estimation of sensitivity and specificity at different cut-off points to identify organ damage in coma patients

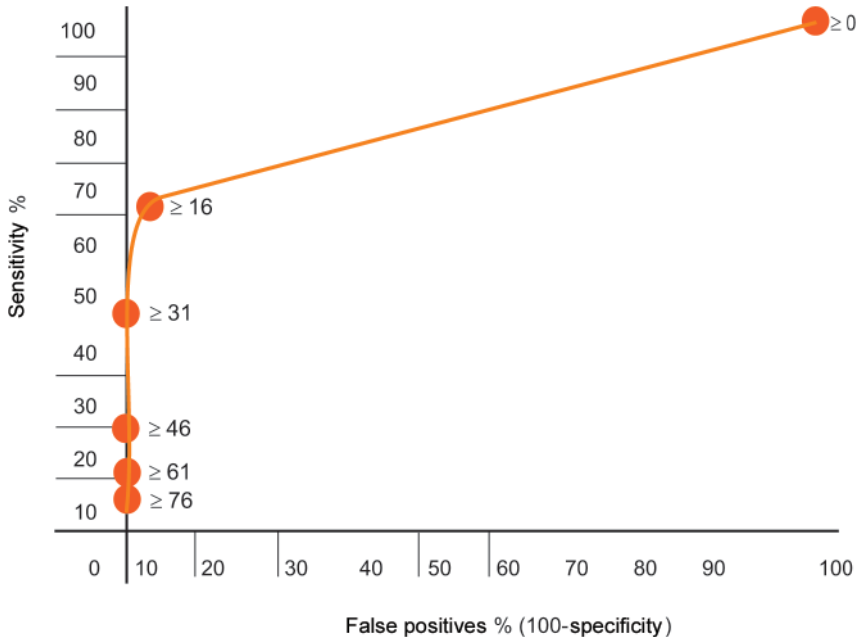


Figure 5 ROC curve

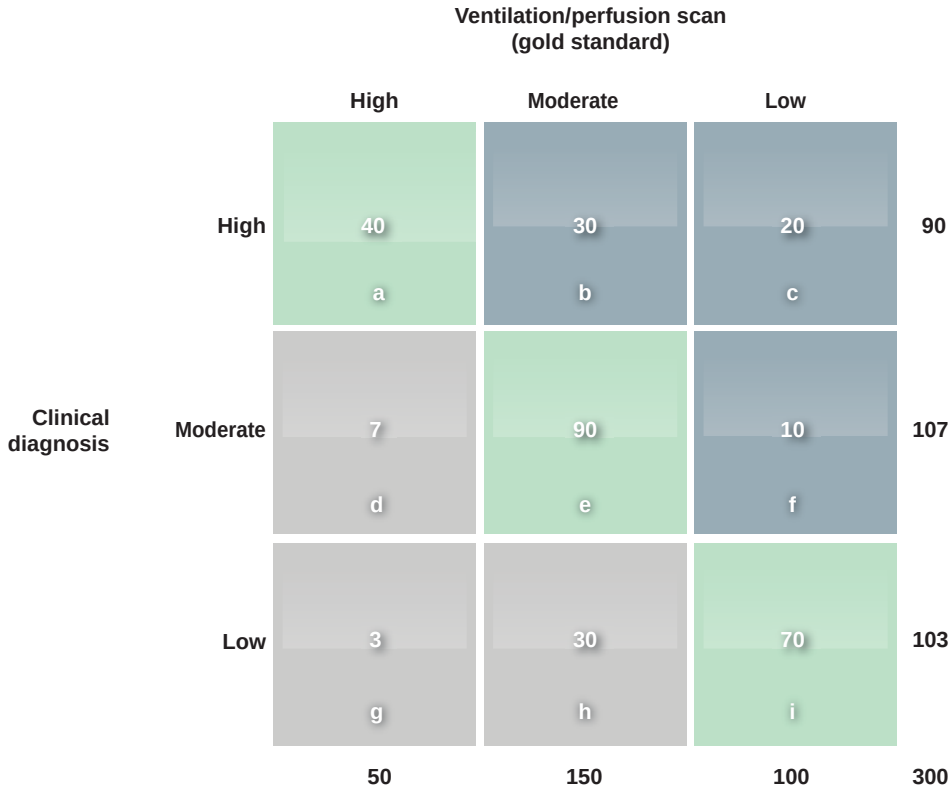


Figure 6 Assessment of clinical diagnosis efficacy in identifying pulmonary thromboembolism

Cells *a, e, i* = matches, in this case 66.66 %
Cells *b, c, f* = errors overestimating the diagnosis, in this case 20 %
Cells *d, g, h* = errors underestimating the diagnosis, in this case 13.33 %

Ordering tests in excess, whether justified or not, generates abnormal results even in normal people, which in turn triggers a cascade of more expensive and riskier tests, in addition with anxiety for the patient.

Common Errors When Elaborating a Diagnostic Test

We already explained how to estimate the efficacy of a diagnostic test and how to make use of it; however, we should watch out for possible causes of systematic errors, with two of them standing out in particular:

1. Inadequate selection of patients.
2. Inadequate interpretation of both the test under study and the gold standard.

The selection of an inadequate spectrum of patients may happen from the clinical or the pathological point of view. For example: the efficacy of a sputum cytology study is not the same for the detection of lung cancer in a patient with a history of heavy and prolonged smoking, weight loss, cough with hemoptysis and dyspnea, than in a patient who only has a cough and whitish expectoration, nor is the efficacy of carcinoembryonic antigen measurement equal for the detection of colon cancer in a patient with Dukes' stage A, compared with a patient with stage D. It is essential for every diagnostic test to be performed with the participation of patients that cover the entire spectrum of the disease, and, in addition, that the proportion of patients in each stratum is reported, so that its usefulness in other populations can be determined. On the other hand, concomitant diseases and used therapies that may alter the efficacy of the test under study should be considered. The control group must have been selected with the same criteria than the problem group, i.e., using the same entrance door, in order for the comparison to have clinico-methodological significance.

With regard to the most common mistakes during the elaboration of a diagnostic test, it is common that when assessing the test under study, the result for the gold standard is already known; this generates an interpretation bias because the assessor is expecting a certain result. Occasionally, the performance and the assessment of the test under study precede the gold

standard and influence on the selection of patients undergoing the latter, or on its interpretation when it has a subjective component and, not infrequently, the test under study is part of the gold standard with which it is compared. All these deviations overestimate the usefulness of the test.

These two large errors can be avoided during the execution of a diagnostic test if the sensitivity and specifity values are considered only when:

a) The spectrum of the disease in the population where it is to be applied is equal to the spectrum of the disease with which the study was developed.

b) The assessment of the test under study and the gold standard has been performed in a blinded and independent manner in all patients.

Finally, it should be emphasized that if the quality of a diagnostic test depends partially on mathematical strategies, the clinical judgment that it derives from is more relevant. And although the sensitivity and specificity estimation starts with the presence or not of the disease, in clinical practice, the study of the patient occurs with the presence or absence of the symptom or sign (clinical or para-clinical).

Additionally, in all cases, the reproducibility of the test should be assessed, provided that the groups under study are comparable; this means that, in addition to the selection of both populations under the same criteria, the distribution of subjects within the different degrees of disease severity must be similar. It should be remembered that, in everyday practice, patients are treated one at a time and that, therefore, it is essential to have a full knowledge of the severity of the disease in the group under study for its subsequent application, so that the patient can be assessed and treated according to the severity of his/her condition and not according to the average severity of the disease in the group in which the diagnostic test or treatment were assessed.

Bibliography

1. Altman DG, Bland JM. Diagnostic tests 1: sensitivity and specificity. BMJ. 1994;308:1552.
2. Altman DG, Bland JM. Diagnostic tests 2: predictive values. BMJ. 1994;309:102.
3. Fagan TJ. Nomogram for Bayes's theorem. N Engl J Med. 1975;293:257.
4. Feinstein AR. Clinical epidemiology. The architecture of clinical research. Philadelphia: W. B. Saunders Company; 1985.
5. Grund B, Sabin C. Analysis of biomarker data: logs, odds ratios, and receiver operating characteristic curves. Curr Opin HIV AIDS. 2010;5(6):473-9.
6. Jaeschke R, Guyatt G, Lijmer J. Diagnostic tests. En: Guyatt G, Rennie D, editors. Users' guides to the medical literature. Chicago: AMA Press; 2002: p. 121-140.
7. Sackett DL, Straus S, Richardson WS, Rosenberg W, Haynes RB. Evidence-based medicine. How to practice and teach EBM. Second edition. Edinburgh: Churchill Living-stone; 2000. p. 67-93.
8. Sackett DL, Haynes RB. The architecture of diagnostic research. BMJ. 2002:324;7336-56.
9. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. Clinical epidemiology. A basic science for clinical medicine. Third edition. US: Little Brown; 2009.
10. Soreide K, Korner H, Soreide JA. Diagnostic accuracy and receiver-operating characteristics curve analysis in surgical research and decision making. Ann Surg. 2011; 253(1):27-34.
11. Tripepi G, Jager KJ, Dekker FW, Zoccali C. Diagnostic methods 2: receiver. operating characteristic (ROC) curves. Kidney Int. 2009;76(3):252-6.